SCALABLE COX REGRESSION VIA GROUPED RISK SETS AND SHARPER LOGSUMEXP RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the computational challenges of large-scale Cox regression, we study stochastic minimization of LogSumExp objectives over large sets. Mini-batch normalizer estimates generally yield biased gradients. We instead use a softplus surrogate that introduces one auxiliary scalar per normalizer and admits unbiased single-sample gradients. For smooth convex LogSumExp objectives, we prove an $O(T^{-1/2})$ averaged objective bound, improving the previous $T^{-1/4}$ analysis. With a strongly convex regularizer on the original variable, we also obtain a last-iterate squared-error rate of $\widetilde{O}(T^{-1})$ without strong convexity in the auxiliary variables. For Cox regression, the normalizers are defined over nested risk sets. We exploit this structure by grouping neighboring failures and sharing one auxiliary variable per group. The resulting compressed objective admits uniform score and curvature bounds that control the errors from grouping and softplus approximation. Together with the general optimization result, these bounds give a mean-square rate of $T^{-4/5}$, up to logarithmic factors, relative to the full Cox solution. The compressed estimator also matches the full estimator's asymptotic distribution. Experiments on synthetic and real survival datasets with slowly decreasing risk sets show a favorable performance relative to stochastic baselines.

1 INTRODUCTION

The motivating problem for this work is large-scale Cox regression. The Cox proportional-hazards model relates covariates to a right-censored event time without specifying the baseline hazard (Cox, 1972). Its partial likelihood associates each failure with a LogSumExp normalizer over everyone still at risk. Nested risk sets admit efficient cumulative-sum and coordinate-descent implementations for linear Cox models (Simon et al., 2011), but these methods rely on full passes over the data. They are less natural when access is stochastic, data are distributed or out of core, or a full pass is itself the resource to avoid.

The Cox bottleneck is an instance of a broader optimization problem: large-set LogSumExp normalizers also arise in numerous applications, including large-class softmax models, distributionally robust optimization, and contrastive learning (Fagan & Iyengar, 2018; Levy et al., 2020; Wei et al., 2026a). Replacing a normalizer by a mini-batch estimate generally biases the gradient of the original objective. Variational formulations instead introduce one auxiliary variable per normalizer and admit unbiased single-sample gradients (Ben-Tal & Teboulle, 1986; Fagan & Iyengar, 2018). Motivated by the Cox problem, we sharpen the analysis of the smooth, tunable softplus formulation of Gladin et al. (2025). For general smooth convex LogSumExp objectives we obtain an $O(T^{-1/2})$ averaged rate. With a strongly convex regularizer on the original parameter, we also obtain an $O(\log T/T)$ last-iterate squared-distance rate, even though the auxiliary variables are not strongly convex.

The nested Cox structure supplies a second, equally important ingredient. Nearby failure times have similar risk sets, so we group consecutive failures and share one normalizer variable within each group.

A single group–event–subject draw gives an unbiased stochastic gradient of the resulting surrogate. Importantly, the gradient perturbation from grouping is second order in the relative risk-set change: exponential reweighting within a group is multiplied by the variation of neighboring risk-set gradients.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This sharper gradient control permits substantial compression while controlling error relative to the full Cox optimizer and preserving the estimator's first-order statistical behavior.

Our contributions are:

- For smooth convex LogSumExp objectives, we establish an $O(T^{-1/2})$ averaged objective bound for the softplus method. When a strongly convex regularizer on the original parameter is added, we further obtain an $O(\log T/T)$ last-iterate squared-distance bound without assuming strong convexity in the auxiliary variables.
- For Cox regression, we give a deterministic compression theorem that controls the gradient and Hessian uniformly through grouping-independent measures of risk-weight dispersion. The gradient error caused by grouping is second order in the within-group risk-set change, while the softplus error is first order in its approximation level. These bounds transfer curvature from the full Cox objective and, together with the general optimization result, yield mean squared parameter error $O((\log T/T)^{4/5})$ relative to the full Cox solution.
- We prove that the compressed and full Cox estimators have the same first-order limit under vanishing approximation schedules, while using a sublinear number of auxiliary shifts. Experiments on synthetic and real survival data evaluate the resulting accuracy–computation tradeoff against stochastic Cox baselines.

The general optimization results apply to smooth convex LogSumExp objectives, our survival-analysis claims concern the convex linear Cox model. This scope lets us separate computational approximation from statistical model error and gives proofs that can be checked directly. Section 3 develops the general optimization guarantees. Section 4 specializes the construction to Cox regression, establishes the deterministic approximation properties, derives the end-to-end computational rate, and gives the statistical transfer result. The appendix contains all auxiliary lemmas and complete proofs.

2 Related work

Scalable Cox optimization. Exact Cox solvers exploit nested risk sets and sparsity but require full-data passes (Simon et al., 2011). Achab et al. (2015) approximate each risk-set expectation by MCMC inside a variance-reduced method. Their strongly convex guarantee uses increasing MCMC accuracy and periodic full gradients. Unbiased multilevel gradients provide an exact stochastic-composition route (Blanchet et al., 2017). Optimal subsampling and one-step correction instead approximate or recover the full-data estimator through a reduced subject sample (Zhang et al., 2024; Wang et al., 2024).

Classical nested case-control estimators have large-sample guarantees under correct Cox specification (Goldstein & Langholz, 1992). Sampled-risk-set neural losses adapt this construction (Kvamme et al., 2019), while online small-stratum objectives (Tarkhan & Simon, 2024) and ordinary mini-batch partial likelihood (Zeng et al., 2026) target related stochastic criteria. In particular, the mini-batch estimator generally optimizes a batch-size-dependent population objective rather than the realized full-data partial likelihood. We instead retain a deterministic, uniformly controlled approximation to the fixed full-data Cox gradient while compressing its outer family of normalizers. Learned normalizer predictors can also compress per-example state (Wei et al., 2026a); our construction uses nested risk sets and provides uniform gradient and Hessian guarantees.

LogSumExp optimization. The Cox loss is a finite sum of coupled compositional objectives of the type studied by Wang & Yang (2022). Single-loop methods obtain near-optimal rates when the outer functions are convex and nondecreasing (Wang & Yang, 2025), whereas the direct logarithmic decomposition falls outside that setting. For the softplus formulation, where $\rho > 0$ is the approximation parameter and smaller values give a sharper approximation, Gladin et al. (2025) obtain $O((\rho\sqrt{T})^{-1}+\rho)$, which optimizes to $T^{-1/4}$; our expected-smoothness analysis gives $O((\rho T)^{-1}+\rho)$ and hence $O(T^{-1/2})$.

SCENT optimizes the exact exponential variational form and also obtains an $O(T^{-1/2})$ averaged convex-objective rate (Wei et al., 2026b). Its general guarantee depends on variance quantities accumulated along the optimization trajectory, and therefore does not directly provide an a priori bound with explicit constants. In contrast, our bounds are expressed in terms of fixed problem

parameters. Our strongly convex result additionally gives an $O(\log T/T)$ last-iterate squared-distance rate without uniform strong convexity in the auxiliary directions. Applied directly to Cox, variational methods retain one shift per event; risk-set grouping reduces this state from m to K and quantifies the resulting error relative to the full Cox target.

3 Stochastic optimization of LogSumExp objectives

We first state the optimization results independently of Cox regression. Consider K distributions μ_k with weights $p_k > 0$, $\sum_{k=1}^K p_k = 1$, and write $\mathbb{E}_k$ for expectation with respect to $X \sim \mu_k$. The convex objective under consideration is

$$J(\theta) = R(\theta) + \sum_{k=1}^{K} p_k \log \mathbb{E}_k e^{L_k(X,\theta)} \tag{1}$$

on a compact convex set Θ. Uniformly over the domain, assume that R and every L_k are twice continuously differentiable, $|L_k| \le B$, $\|\nabla L_k\| \le G$, $0 \preceq \nabla^2 L_k \preceq HI$, $\|\nabla R\| \le G_R$, and $0 \preceq \nabla^2 R \preceq L_R I$. Define the normalized exponential moment

$$\bar{\kappa} = \max_k \sup_{\theta\in\Theta} \frac{\mathbb{E}_k e^{2L_k(X,\theta)}}{(\mathbb{E}_k e^{L_k(X,\theta)})^2}.$$

For $0 < \rho < 1$, let $h_\rho(u) = \rho^{-1}\log(1+\rho e^u)$. Building on the softplus construction of Gladin et al. (2025), we introduce the joint surrogate and its profiled counterpart:

$$G_\rho(\theta, s) = R(\theta) + \sum_{k=1}^{K} p_k\{s_k - 1 + \mathbb{E}_k h_\rho(L_k(X,\theta) - s_k)\}, \qquad J_\rho(\theta) = \min_{s\in\mathbb{R}^K} G_\rho(\theta, s). \tag{2}$$

Proposition A.1, based on the approximation bound of Gladin et al. (2025), gives $J + \rho/2 + \log(1-\rho\bar{\kappa}) \le J_\rho \le J$ whenever $\rho\bar{\kappa} < 1$. If $\rho\bar{\kappa} \le 1/8$, all optimal shifts lie in $[-B-1, B]^K$.

For $z = (\theta, s)$ and a gradient $g = (g_\theta, g_s)$, let $P = \mathrm{diag}(I_d, p_1, \dots, p_K)$ and use the matrix-induced norm and its dual:

$$\|z\|_P^2 = z^\top P z = \|\theta\|^2 + \sum_k p_k s_k^2, \qquad \|g\|_{P^{-1}}^2 = g^\top P^{-1} g = \|g_\theta\|^2 + \sum_k \frac{g_{s,k}^2}{p_k}.$$

At each iteration, sample k uniformly from $\{1, \dots, K\}$ and X from its group distribution. With

$$w = \frac{1}{\rho + \exp\{s_k - L_k(X,\theta)\}},$$

an unbiased stochastic gradient of G_ρ is

$$g_\theta = \nabla R(\theta) + K p_k w \nabla L_k(X,\theta), \qquad g_{s,k} = K p_k (1-w),$$

with all other shift coordinates zero. We apply the corresponding weighted projected step on $\mathcal{Z} := \Theta \times [-B-1, B]^K$:

$$z_{t+1} = \arg\min_{z\in\mathcal{Z}} \left\{ \eta \langle g_t, z\rangle + \frac{1}{2}\|z - z_t\|_P^2 \right\}. \tag{3}$$

Equivalently, the parameter uses an ordinary projected step, while the sampled shift is clipped after $s_k \leftarrow s_k - \eta K(1-w)$; all other shifts stay fixed. Initialize z_1 anywhere in this domain. Let $z_\rho = (\theta_\rho, s_\rho)$ minimize G_ρ there and define

$$C = L_R + KH + \frac{K}{4}(G^2+1), \qquad V = 2G_R^2 + 2K + 8K\bar{\kappa}(G^2+1), \tag{4}$$

$$D^2 = \mathrm{diam}(\Theta)^2 + (2B+1)^2. \tag{5}$$

The next two results cover the general convex regime with iterate averaging and the strongly convex regime with last-iterate convergence, respectively.

Theorem 3.1. *Set $\rho_T = T^{-1/2}$ and $\eta_T = (4C\sqrt{T})^{-1}$, and suppose $T \geq 64\bar{\kappa}^2$. For the averaged parameter $\bar{\theta}_T = T^{-1}\sum_{t=1}^{T}\theta_t$,*

$$\mathbb{E}[J(\bar{\theta}_T) - \min_{\Theta} J] \leq \frac{4CD^2}{\sqrt{T}} + \frac{V}{2C\sqrt{T}} + \frac{8\bar{\kappa}}{7\sqrt{T}}. \tag{6}$$

Thus the softplus method attains an $O(T^{-1/2})$ objective rate for the original LogSumExp problem.

Theorem 3.2. *Assume $0 < \rho < 1$, $\rho\bar{\kappa} \leq 1/8$, and that J_ρ is γ-strongly convex. Define*

$$\mu = \left\{\frac{2(1+2G^2)}{\gamma} + 16B + 8\right\}^{-1}. \tag{7}$$

For a constant step size $0 < \eta \leq \min\{\rho/(4C), 1/\mu\}$, the last iterate satisfies

$$\mathbb{E}\,\|z_{T+1} - z_\rho\|_P^2 \leq e^{-\mu\eta T}D^2 + \frac{2\eta V}{\mu}. \tag{8}$$

In particular, for $T \geq 2$, set $\rho_T = 8C\log T/(\mu T)$ and $\eta_T = \rho_T/(4C)$. If these choices satisfy the preceding conditions, then

$$\mathbb{E}\,\|\theta_{T+1} - \theta_{\rho_T}\|^2 \leq \frac{D^2}{T^2} + \frac{4V\log T}{\mu^2 T}. \tag{9}$$

Corollary 3.3. *Let $\theta^\star$ minimize J and assume that it is stationary, and define $L_J = L_R + H + G^2$. Under the scheduled choices of Theorem 3.2, the last iterate satisfies*

$$\mathbb{E}[J(\theta_{T+1}) - J(\theta^\star)] \leq \frac{L_J D^2}{T^2} + \frac{4L_J V\log T}{\mu^2 T} + \frac{16L_J\rho_T\bar{\kappa}}{7\gamma} = O\left(\frac{\log T}{T}\right). \tag{10}$$

If R is λ-strongly convex, then so is J_ρ, and one may take $\gamma = \lambda$. Appendix A proves the two theorems and the corollary using the expected-smoothness estimate of Lemma A.2 and, under strong convexity, the restricted-secant inequality of Lemma A.4.

4 Cox Regression with Grouped Risk Sets

We now introduce grouped Cox risk sets and specialize the construction of Section 3 to the resulting objective.

4.1 Cox Objective and Risk-Set Grouping

Full Cox objective. Consider N subjects. Subject j has observed time y_j, event indicator $\Delta_j \in \{0,1\}$, and covariate vector $x_j \in \mathbb{R}^d$. For simplicity, we assume distinct observed failure times.[1] Let $\mathcal{E} = \{j : \Delta_j = 1\}$ and assume $m = |\mathcal{E}| \geq 1$. For $i \in \mathcal{E}$, define the risk set and its size by

$$\mathcal{R}_i = \{j : y_j \geq y_i\}, \qquad n_i = |\mathcal{R}_i|.$$

For a coefficient vector β in a compact convex set $\mathcal{B} \subset \mathbb{R}^d$, write

$$a_i(\beta) = \log\left(\frac{1}{n_i}\sum_{j\in\mathcal{R}_i} e^{\beta^\top x_j}\right), \qquad \mathcal{L}(\beta) = \frac{\lambda}{2}\|\beta\|^2 - \beta^\top\bar{x}_{\mathcal{E}} + \frac{1}{m}\sum_{i\in\mathcal{E}} a_i(\beta), \tag{11}$$

where $\lambda \geq 0$ is an optional ridge coefficient and $\bar{x}_{\mathcal{E}} = m^{-1}\sum_{i\in\mathcal{E}} x_i$. The factors $1/n_i$ add only a constant to the usual negative log partial likelihood, so they do not change its minimizer for a fixed λ (Cox, 1972). Setting $\lambda = 0$ recovers the unpenalized objective. We call $\beta^\top x_j$ the linear predictor and $e^{\beta^\top x_j}$ the risk weight of subject j.

[1] The finite-data compression results accommodate tied failures under the Breslow convention (Breslow, 1974), as does the experimental implementation (Appendix D). The statistical transfer result assumes distinct failures.

Compressing neighboring risk sets. Order failures by time and greedily partition them into consecutive groups $\mathcal{I}_1, \ldots, \mathcal{I}_K$. A group is extended for as long as

$$\frac{\max_{i \in \mathcal{I}_k} n_i}{\min_{i \in \mathcal{I}_k} n_i} \leq 1 + \delta, \tag{12}$$

where $\delta > 0$ is chosen by the user. Equivalently, setting $r_\delta := \delta/(1+\delta)$ bounds the fraction of the largest risk set that disappears within a group by r_δ. Let $m_k = |\mathcal{I}_k|$ and $p_k = m_k/m$. Conditional on group k, draw an event I uniformly from $\mathcal{I}_k$ and then a subject J uniformly from $\mathcal{R}_I$. We use $\mathbb{E}_k$ for this two-stage expectation and set

$$\Phi_k(\beta) = \log \mathbb{E}_k e^{\beta^\top x_J} = \log \left(\frac{1}{m_k} \sum_{i \in \mathcal{I}_k} e^{a_i(\beta)} \right). \tag{13}$$

Replacing the average of the a_i in each group by Φ_k gives the grouped objective $\widetilde{\mathcal{L}}_\delta(\beta) = \lambda \|\beta\|^2/2 - \beta^\top \bar{x}_\mathcal{E} + \sum_k p_k \Phi_k(\beta)$. The replacement is exact for singleton groups. More importantly, its gradient error is quadratic in the within-group risk-set change; this is formalized in Theorem 4.1.

For $0 < \rho < 1$, use the softplus function h_ρ from Section 3. Introduce one shift $s_k \in \mathbb{R}$ per group and the joint objective

$$G_{\rho,\delta}(\beta, s) = \frac{\lambda}{2} \|\beta\|^2 - \beta^\top \bar{x}_\mathcal{E} + \sum_{k=1}^{K} p_k \left[s_k - 1 + \mathbb{E}_k h_\rho(\beta^\top x_J - s_k) \right]. \tag{14}$$

Its profiled version is

$$\mathcal{L}_{\rho,\delta}(\beta) = \min_{s \in \mathbb{R}^K} G_{\rho,\delta}(\beta, s). \tag{15}$$

Thus $\mathcal{L}$, $\widetilde{\mathcal{L}}_\delta$, and $\mathcal{L}_{\rho,\delta}$ denote the exact, grouped, and grouped-softplus objectives, all with the same λ. With $R(\beta) = \lambda \|\beta\|^2/2 - \beta^\top \bar{x}_\mathcal{E}$, $L_k((I,J),\beta) = \beta^\top x_J$, and $\Theta = \mathcal{B}$, this is exactly the generic construction in (2) for the two-stage group distributions defined above. The scalar minimizers are unique, finite, and have strictly positive second derivatives. Consequently the envelope and implicit-function theorems justify differentiating the profile and taking the Schur complement of the joint Hessian; details appear in Appendix A.

Stochastic updates and cost model. Assume $\|x_j\| \leq X$ and $|\beta^\top x_j| \leq M$ for every subject and $\beta \in \mathcal{B}$. At iteration t, draw $k_t \sim \mathrm{Unif}\{1, \ldots, K\}$, then $i_t \sim \mathrm{Unif}(\mathcal{I}_{k_t})$, and finally $j_t \sim \mathrm{Unif}(\mathcal{R}_{i_t})$. With

$$\widehat{x}_{\mathcal{E},t} = K p_{k_t} x_{i_t}, \qquad w_t = \frac{1}{\rho + \exp(s_{k_t,t} - \beta_t^\top x_{j_t})},$$

we have $\mathbb{E}_t \widehat{x}_{\mathcal{E},t} = \bar{x}_\mathcal{E}$. One projected stochastic-gradient step is

$$\beta_{t+1} = \mathrm{proj}_\mathcal{B}[\beta_t - \eta\{\lambda\beta_t + K p_{k_t}(w_t x_{j_t} - x_{i_t})\}], \tag{16}$$

$$s_{k_t,t+1} = \mathrm{clip}_{[-M-1,M]}[s_{k_t,t} - \eta K(1 - w_t)], \tag{17}$$

and all unsampled shifts stay fixed. The sampled Euclidean gradient of (14) has coefficient component $\lambda\beta_t + K p_{k_t}(w_t x_{j_t} - x_{i_t})$ and shift component $g_{s,k_t} = K p_{k_t}(1 - w_t)$, with all other shift components zero. This gradient estimator is unbiased. The weighted update divides the sampled shift component by p_{k_t}, giving (17); see Appendix A.

After event times have been sorted and group boundaries stored, an iteration uses two sampled covariate vectors and $O(d)$ arithmetic. The optimizer state is $O(d + K)$ beyond storage or streaming access to the data. We assume constant time access to a uniformly sampled member of an indexed risk set and an inexpensive projection onto $\mathcal{B}$.

4.2 Gradient and Curvature Preservation

We first introduce the two data quantities that govern the result. The normalized risk-weight second moment is

$$\kappa = \sup_{\beta \in \mathcal{B}} \max_{i \in \mathcal{E}} \frac{n_i^{-1} \sum_{j \in \mathcal{R}_i} e^{2\beta^\top x_j}}{(n_i^{-1} \sum_{j \in \mathcal{R}_i} e^{\beta^\top x_j})^2}. \tag{18}$$

Fix $q \in (0,1)$. To control the risk weights of subjects leaving over a short time interval, define

$$L_q = 1 \vee \sup_{\beta \in \mathcal{B}} \max_{u,v \in \mathcal{E},\ \mathcal{R}_v \subsetneq \mathcal{R}_u,\ n_v \geq q n_u} \frac{|\mathcal{R}_u \setminus \mathcal{R}_v|^{-1} \sum_{j \in \mathcal{R}_u \setminus \mathcal{R}_v} e^{\beta^\top x_j}}{n_u^{-1} \sum_{j \in \mathcal{R}_u} e^{\beta^\top x_j}}. \tag{19}$$

For fixed β and $\mathcal{R}_u$, the maximum ranges over later nested risk sets $\mathcal{R}_v$ that retain at least a q fraction of $\mathcal{R}_u$; the ratio compares the mean risk weight of the subjects who leave between the two event times with the mean risk weight in $\mathcal{R}_u$. If the maximum has no admissible pair, it is omitted. Both quantities are defined independently of the grouping. They are finite under bounded linear predictors, but keeping them explicit replaces a worst-case exponential range factor by the observed normalized dispersion and local risk-weight ratio.

Theorem 4.1. *Suppose $r_\delta \leq 1 - q$, $L_q r_\delta \leq 1/2$, and $\rho\kappa \leq 1/9$. Then, uniformly for $\beta \in \mathcal{B}$,*

$$\|\nabla \mathcal{L}_{\rho,\delta}(\beta) - \nabla \mathcal{L}(\beta)\| \leq X\{L_q^2 r_\delta^2 + 4\rho\kappa\}, \tag{20}$$

$$\nabla^2 \mathcal{L}_{\rho,\delta}(\beta) \succeq \nabla^2 \mathcal{L}(\beta) - X^2 \left\{ \frac{5}{2} L_q^2 r_\delta^2 + 20\rho\kappa \right\} I. \tag{21}$$

If $\nabla^2 \mathcal{L}(\beta) \succeq \nu I$ on $\mathcal{B}$ for some $\nu > 0$, and if $\beta^\star$ and $\beta_{\rho,\delta}$ minimize $\mathcal{L}$ and $\mathcal{L}_{\rho,\delta}$ on $\mathcal{B}$, respectively, then

$$\|\beta_{\rho,\delta} - \beta^\star\| \leq \frac{X}{\nu}\{L_q^2 r_\delta^2 + 4\rho\kappa\}. \tag{22}$$

For $\lambda > 0$, the ridge term supplies the full-objective curvature assumption with $\nu = \lambda$; when $\lambda = 0$, this assumption depends on the data and design. The Hessian bound also shows that if the matrix error subtracted in (21) is at most $\nu/2$, then $\mathcal{L}_{\rho,\delta}$ is $\nu/2$-strongly convex. The quadratic grouping error comes from multiplying the $O(L_q r_\delta)$ change in event weights by the $O(X L_q r_\delta)$ variation of event gradients within a group. Softplus profiling contributes $O(\rho\kappa)$, and the Hessian comparison retains the nonnegative covariance term. Appendix B proves the gradient, curvature, and minimizer bounds.

4.3 End-to-end computational rate

We now combine the generic last-iterate guarantee of Theorem 3.2 with the Cox gradient and curvature bounds in Theorem 4.1.

Corollary 4.2. *Assume $\nabla^2 \mathcal{L}(\beta) \succeq \nu I$ on $\mathcal{B}$ for some $\nu > 0$, fix $q \in (0,1)$, and suppose every event risk set contains at least cN subjects for a fixed $c \in (0,1)$. Let the problem constants be fixed as T varies. For all sufficiently large declared horizons T, use maximal grouping and the schedule*

$$\delta_T = \left(\frac{\log T}{T}\right)^{1/5}, \qquad \rho_T = \frac{8 A_{K_T} \log T}{\mu T}, \qquad \eta_T = \frac{2 \log T}{\mu T}, \tag{23}$$

where $A_K = \lambda + K(X^2+1)/4$ and $\mu = \{4(1+2X^2)/\nu + 16M + 8\}^{-1}$. Initialize (β_1, s_1) in $\mathcal{B} \times [-M-1, M]^{K_T}$. Then, conditional on the data, the last iterate of (16)–(17) satisfies

$$\mathbb{E}\,\|\beta_{T+1} - \beta^\star\|^2 = O\left((\log T/T)^{4/5}\right). \tag{24}$$

When $\beta^\star$ is interior, the exact-objective error has the same order, and the auxiliary state is $K_T = O((T/\log T)^{1/5})$ until singleton groups cause finite-data saturation.

Indeed, maximal grouping gives $K_T = O(\delta_T^{-1})$, so the stochastic error is $O(\log T/(\delta_T T))$, whereas the squared grouping bias is $O(\delta_T^4)$. Balancing these terms gives (24); Appendix B gives the exact finite-time bound and all smallness conditions.

4.4 First-order statistical efficiency

We now let the data set vary with sample size N and set $\lambda = 0$ in all objectives on a fixed compact convex set $\mathcal{B}$. Fix $q \in (0,1)$ and assume $\max_{1 \leq j \leq N} \|x_j\| \leq X$ uniformly in N. Let $\widehat{\beta}_N$ minimize the full empirical objective $\mathcal{L}_N$, and let $\widehat{\beta}_N^{\rho,\delta}$ minimize its grouped-softplus counterpart. Classical conditions for the Cox model yield asymptotic normality of the maximum partial-likelihood estimator (Tsiatis, 1981; Andersen & Gill, 1982). We take the asymptotic normality of the full Cox estimator as given and isolate the additional conditions under which it transfers to the grouped-softplus estimator.

Theorem 4.3. *Suppose $\sqrt{N}(\widehat{\beta}_N - \beta_0) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\beta_0)^{-1})$ for a positive-definite information matrix $\mathcal{I}(\beta_0)$. Assume that, with probability tending to one, $\widehat{\beta}_N$ and $\widehat{\beta}_N^{\rho,\delta}$ lie in a fixed convex neighborhood $U \subset \mathcal{B}$ of β_0 and $\nabla^2 \mathcal{L}_N(\beta) \succeq \nu_0 I$ on U for a fixed $\nu_0 > 0$. Let $L_{q,N}$ and κ_N denote (19) and (18) for the N-subject data set. If*

$$\sqrt{N}\{L_{q,N}^2 \delta_N^2 + \rho_N \kappa_N\} \xrightarrow{p} 0, \tag{25}$$

then

$$\sqrt{N}\left\|\widehat{\beta}_N^{\rho,\delta} - \widehat{\beta}_N\right\| \xrightarrow{p} 0, \qquad \sqrt{N}(\widehat{\beta}_N^{\rho,\delta} - \beta_0) \Longrightarrow \mathcal{N}(0, \mathcal{I}(\beta_0)^{-1}).$$

In particular, if $L_{q,N} = O_p(1)$ and $\kappa_N = O_p(1)$, then for any $\ell_N \to \infty$ one may choose

$$\delta_N = N^{-1/4}/\ell_N, \qquad \rho_N = N^{-1/2}/\ell_N.$$

If every event risk set contains at least cN subjects with probability tending to one for some fixed $c \in (0, 1)$, and if $\ell_N = o(N^{3/4})$, maximal grouping uses $K_N = O_p(N^{1/4}\ell_N) = o_p(N)$ shifts.

The proof in Appendix C combines the uniform gradient bound in Theorem 4.1 with local strong convexity and Slutsky's theorem.

Numerical accuracy. Suppose that the assumptions of Corollary 4.2 hold with constants uniform in N: the curvature and minimum risk-set fraction are bounded away from zero, while X, M, $L_{q,N}$, and κ_N are bounded above. Then a run of T_N iterations has the same first-order limit as the full empirical Cox estimator whenever

$$N\left(\frac{\log T_N}{T_N}\right)^{4/5} \longrightarrow 0, \tag{26}$$

by Markov's inequality and (24). The precise joint numerical and statistical conditions are recorded in Appendix C.

5 Experiments

5.1 Benchmark protocol

We compare our method with Batch LSE, Minibatch Cox (Zeng et al., 2026), BigSurv (Tarkhan & Simon, 2020), and Cox-CC (Kvamme et al., 2019) on SUPPORT2 (Vanderbilt Biostatistics, n.d.), NWTCO (R survival package documentation, n.d.), and three synthetic datasets (Table 1). All methods share a stratified 60/20/20 train/validation/test split on each dataset, with preprocessing fitted on training data and Breslow handling of ties (Breslow, 1974).

Table 1: Datasets, training event counts m_{tr}, group counts K at $\delta = 0.05$, and per-fit vector budgets (millions). Dimension d is after preprocessing.

Dataset	N	d	m_{tr}	K	Tuning cap	Final cap
SUPPORT2	8,873	22	3,621	61	3	4.5
NWTCO	4,028	11	342	15	3	4.5
Correlated 100k	100,000	10	6,005	133	10	15
Correlated 1M	1,000,000	20	60,136	178	100	150
Independent 100k	100,000	20	41,921	166	10	15

All methods initialize the coefficient vector at $\beta_0 = 0$ and include the ridge penalty $\lambda\|\beta\|_2^2/2$, with $\lambda = 10^{-3}$. Our method averages mini-batches of b independent stochastic gradient samples; although the theory is stated for one oracle sample per iteration, such averaging preserves unbiasedness and the moment bounds used in the analysis. We use the decreasing step size $\eta_t = \eta_0(1 + t/1000)^{-1/2}$ and initialize the auxiliary shifts at $s_{k,0} = \log(1 - \rho)$. We fix $\rho = 10^{-4}$ and $\delta = 0.05$. The batch size b and initial step size η_0 are selected by validation. At each saved checkpoint after initialization, optimization and validation curves for our method are evaluated at $\bar{\beta}_t = t^{-1}\sum_{r=1}^{t}\beta_r$, the arithmetic mean of the projected post-update iterates.

For each method and dataset, we evaluate twelve configurations with two tuning seeds. We select the configuration with the smallest mean of the two seeds' minimum saved validation Cox losses

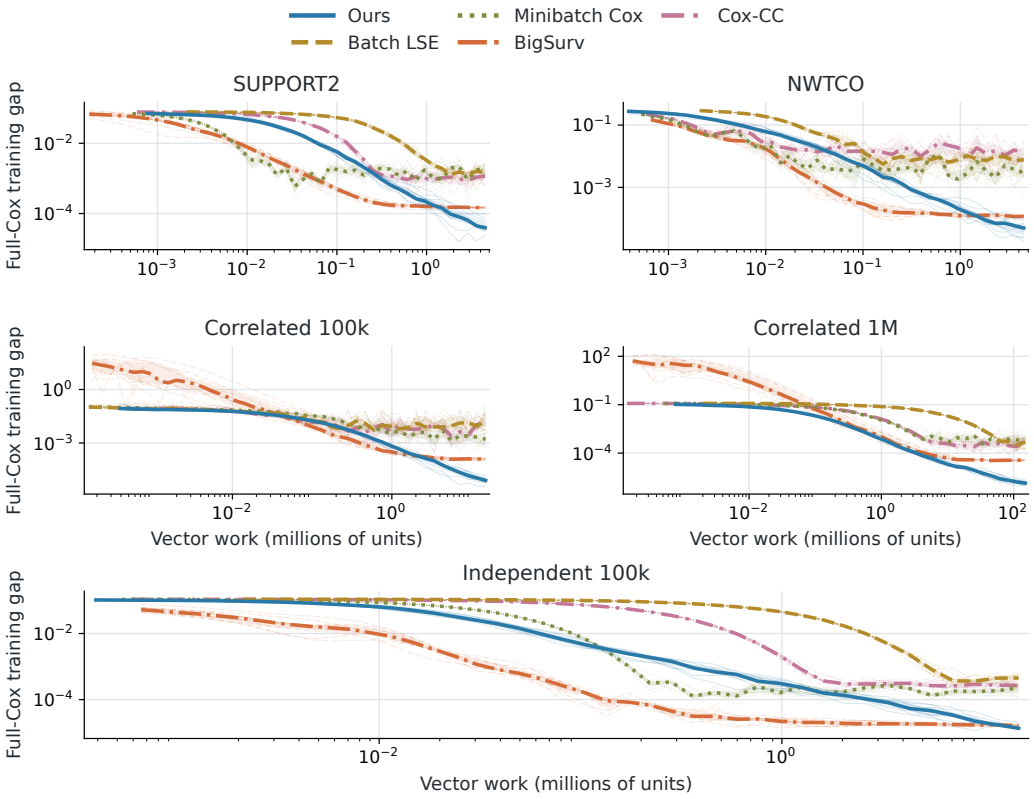


Figure 1: Regularized full-Cox training gap against vector work. Each dataset has the same cap for all five methods. Thin lines show ten optimizer seeds, thick lines their medians, and bands their interquartile ranges. Ours and BigSurv use running arithmetic averages of post-update coefficient iterates. The other methods use current coefficient iterates.

and run it with ten optimizer seeds. All methods receive the same per-fit vector budgets: $U_{\mathrm{HPO}} = \max\{3 \cdot 10^6, 100N\}$ for tuning and $U_{\mathrm{final}} = 1.5U_{\mathrm{HPO}}$ for final runs, where N is the total number of subjects.

We measure training accuracy by the regularized full-Cox gap $\mathcal{L}(\beta) - \mathcal{L}(\beta_{\mathrm{ref}})$ to a numerical reference. Test Cox loss and Harrell's C-index are evaluated at validation-selected checkpoints; predictive Cox loss is unpenalized. Appendix D details the baseline objectives, vector-work accounting, preprocessing, and selection and evaluation history.

5.2 Optimization and Prediction

Our method achieves the lowest median terminal regularized full-Cox training gap on all five datasets at matched vector-work budgets (Figure 1). BigSurv is the closest baseline, with median terminal gaps 1.22–26.15 times ours (Table 4).

At $\delta = 0.05$, grouping reduces the number of auxiliary shifts by factors of 22.8–337.8 relative to one shift per training event (Table 1).

Additionally, to isolate the two approximation mechanisms in Theorem 4.1, we measure at the full-Cox reference β_{ref}

$$e_{\mathrm{group}}(\delta) = \|\nabla\widetilde{\mathcal{L}}_\delta(\beta_{\mathrm{ref}}) - \nabla\mathcal{L}(\beta_{\mathrm{ref}})\|, \qquad e_{\mathrm{soft}}(\rho) = \|\nabla\mathcal{L}_{\rho,0.05}(\beta_{\mathrm{ref}}) - \nabla\widetilde{\mathcal{L}}_{0.05}(\beta_{\mathrm{ref}})\|.$$

Figure 2 examines the quadratic grouping and linear softplus terms in (20). Softplus error is nearly linear in ρ across all datasets. The grouping slopes are 1.37–1.87; because the partition changes discretely with δ, the $O(\delta^2)$ bound does not imply an exact finite-grid slope of two.

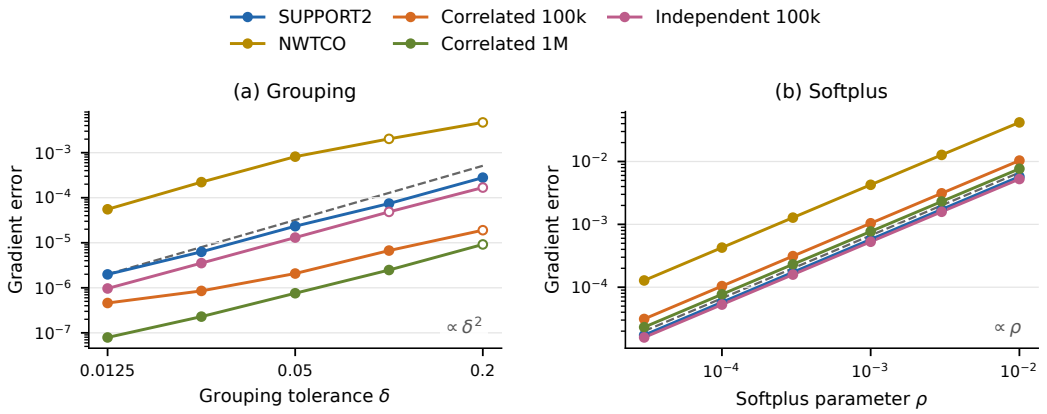


Figure 2: Coefficient-gradient approximation errors at β_{ref}. The left panel isolates risk-set grouping using exact group normalizers; the right isolates softplus profiling at $\delta = 0.05$. Filled markers satisfy the measured pointwise conditions of Theorem 4.1; hollow markers violate at least one applicable condition. Dashed lines are proportionality guides with arbitrary vertical offsets.

Our method has the lowest mean test Cox loss on NWTCO, Correlated 100k, and Independent 100k. Minibatch Cox has the lowest mean on SUPPORT2, and BigSurv on Correlated 1M. Full metrics and trajectories are reported in Appendices E and G.

Appendix F gives the parameter grids, diagnostics at additional coefficient vectors, fitted slopes, numerical profiling details, and pointwise checks of the sufficient conditions.

6 Limitations and Conclusion

Risk-set grouping converts the full Cox objective into a stochastic problem with K shared normalizers while preserving its gradient to second order in the grouping tolerance. Under the stated approximation and curvature conditions, this yields the $\widetilde{O}(T^{-4/5})$ computational rate. Under the additional statistical conditions and parameter schedules, the compressed estimator is first-order equivalent to the full Cox estimator, with sublinear auxiliary state under the stated risk-set size condition. The method is intended for stochastic-access regimes; the experiments do not establish an end-to-end runtime advantage over optimized cumulative-sum full-Cox solvers.

The analysis assumes a convex linear predictor and the stated boundedness and curvature conditions. The constants L_q and κ can be large under adverse risk-weight heterogeneity. The implementation handles equal failure times with the Breslow convention, whereas the main statistical transfer result is stated for distinct failures. Tied-time asymptotics, time-varying covariates, and nonlinear predictors remain outside its scope.

At matched vector-work budgets, our method achieves the smallest median terminal regularized full-Cox training gap on all five datasets. Together, the theory and experiments establish a stochastic approach to full-Cox optimization that combines explicit approximation guarantees with compact auxiliary state.

Reproducibility Statement

The appendix provides complete proofs and a detailed account of the experimental setup, implementation, tuning, and evaluation procedures. The accompanying source code is included in the supplementary material.

AI USE STATEMENT

Generative AI assisted with language editing, literature searches, and proof development. All AI-assisted mathematical content was independently verified, and the authors take full responsibility for the manuscript.

REFERENCES

Massil Achab, Agathe Guilloux, Stéphane Gaïffas, and Emmanuel Bacry. SGD with variance reduction beyond empirical risk minimization. *arXiv preprint arXiv:1510.04822*, 2015. doi: 10.48550/arXiv.1510.04822. URL https://arxiv.org/abs/1510.04822.

Per K. Andersen and Richard D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982. doi: 10.1214/aos/1176345976.

Aharon Ben-Tal and Marc Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986. doi: 10.1287/mnsc.32.11.1445.

Jose Blanchet, Donald Goldfarb, Garud Iyengar, Fengpei Li, and Chaoxu Zhou. Unbiased simulation for optimizing stochastic function compositions. *arXiv preprint arXiv:1711.07564*, 2017. doi: 10.48550/arXiv.1711.07564. URL https://arxiv.org/abs/1711.07564.

Norman E. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974. doi: 10.2307/2529620.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Francois Fagan and Garud Iyengar. Unbiased scalable softmax optimization. *arXiv preprint arXiv:1803.08577*, 2018. doi: 10.48550/arXiv.1803.08577. URL https://arxiv.org/abs/1803.08577.

Egor Gladin, Alexey Kroshnin, Jia-Jie Zhu, and Pavel Dvurechensky. Improved stochastic optimization of LogSumExp. *arXiv preprint arXiv:2509.24894*, 2025. doi: 10.48550/arXiv.2509.24894. URL https://arxiv.org/abs/2509.24894.

Larry Goldstein and Bryan Langholz. Asymptotic theory for nested case-control sampling in the Cox regression model. *The Annals of Statistics*, 20(4):1903–1928, 1992. doi: 10.1214/aos/1176348895.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. URL https://www.jmlr.org/papers/v20/18-424.html.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8847–8860, 2020.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/40008b9a5380fcacce3976bf7c08af5b-Abstract.html.

R survival package documentation. nwtco: Data from the national wilm's tumor study. https://stat.ethz.ch/R-manual/R-devel/library/survival/html/nwtco.html, n.d. Dataset documentation; accessed 2026-09-17.

Noah Simon, Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. doi: 10.18637/jss.v039.i05.

Aliasghar Tarkhan and Noah Simon. BigSurvSGD: Big survival data analysis via stochastic gradient descent. *arXiv preprint arXiv:2003.00116*, 2020. URL https://arxiv.org/abs/2003.00116.

Aliasghar Tarkhan and Noah Simon. An online framework for survival analysis: Reframing Cox proportional hazards model for large data sets and neural networks. *Biostatistics*, 25(1):134–153, 2024. doi: 10.1093/biostatistics/kxac039.

Anastasios A. Tsiatis. A large sample study of Cox's regression model. *The Annals of Statistics*, 9(1): 93–108, 1981. doi: 10.1214/aos/1176345335.

Vanderbilt Biostatistics. SUPPORT Datasets. https://hbiostat.org/data/repo/supportdesc, n.d. Dataset documentation; accessed 2026-09-17.

Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23292–23317, 2022. URL https://proceedings.mlr.press/v162/wang22ak.html.

Bokun Wang and Tianbao Yang. A near-optimal single-loop stochastic algorithm for convex finite-sum coupled compositional optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65091–65121, 2025. URL https://proceedings.mlr.press/v267/wang25dw.html.

Jianqiao Wang, Donglin Zeng, and Dan-Yu Lin. Fitting the Cox proportional hazards model to big data. *Biometrics*, 80(1):ujae018, 2024. doi: 10.1093/biomtc/ujae018.

Xiyuan Wei, Chih-Jen Lin, and Tianbao Yang. NeuCLIP: Efficient large-scale CLIP training with neural normalizer optimization. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=WoMMSVZHfP.

Xiyuan Wei, Linli Zhou, Bokun Wang, Chih-Jen Lin, and Tianbao Yang. A geometry-aware efficient algorithm for compositional entropic risk minimization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2602.02877.

Lang Zeng, Weijing Tang, Zhao Ren, and Ying Ding. Mini-batch estimation for deep Cox models: Statistical foundations and practical guidance. *Journal of the American Statistical Association*, 121 (554):988–999, 2026. doi: 10.1080/01621459.2026.2644611.

Haixiang Zhang, Lulu Zuo, HaiYing Wang, and Liuquan Sun. Approximating partial likelihood estimators via optimal subsampling. *Journal of Computational and Graphical Statistics*, 33(1): 276–288, 2024. doi: 10.1080/10618600.2023.2216261.

A GENERIC GROUPED LOGSUMEXP OPTIMIZATION

This section proves Theorems 3.1 and 3.2. It first establishes the approximation, optimum-noise, and profile-to-joint curvature lemmas, then derives the convex averaged bound and the strongly convex last-iterate recursion.

A.1 SOFTPLUS IDENTITIES AND PROFILE REGULARITY

We use the assumptions and objectives of Section 3. Write $F_k(\theta) := \log \mathbb{E}_k e^{L_k(X,\theta)}$ for one group's LogSumExp term. The scalar weight and curvature identities needed below are

$$w_\rho(a) := h'_\rho(a) = \frac{e^a}{1+\rho e^a}, \qquad h''_\rho(a) = w_\rho(a)[1-\rho w_\rho(a)] \in \left(0, \frac{1}{4\rho}\right].$$

For $0 < \rho < 1$, each scalar objective tends to $+\infty$ at both ends. As $s_k \to +\infty$, its leading term is s_k, while as $s_k \to -\infty$, its leading term is $(1-\rho^{-1})s_k$, both tend to $+\infty$. Thus a scalar minimizer exists, and it is unique because the second derivative with respect to s_k is positive. Boundedness of the losses permits differentiation under the expectation. The minimizer is interior and its scalar second derivative is nonzero, so the implicit-function theorem makes the optimal shift differentiable in θ, the envelope theorem therefore justifies the profile gradients used below.

Proposition A.1. *Suppose $\rho\bar{\kappa} < 1$. Then, for every $\theta \in \Theta$,*

$$J(\theta) + \frac{\rho}{2} + \log(1-\rho\bar{\kappa}) \le J_\rho(\theta) \le J(\theta). \tag{27}$$

If $s_{k,\rho}(\theta)$ is the optimal shift for group k, then

$$F_k(\theta) + \log(1-\rho\bar{\kappa}) \le s_{k,\rho}(\theta) < F_k(\theta). \tag{28}$$

Proof. For each group, apply the softplus approximation and optimal-shift bounds of Gladin et al. (2025) with $\varphi = L_k(X,\theta)$. Their normalized second moment is $\mathbb{E}_k e^{2[L_k - F_k]} \le \bar{\kappa}$, so the groupwise profile lies between $F_k + \rho/2 + \log(1-\rho\bar{\kappa})$ and F_k, and its optimal shift lies in the interval (28). Weighting the value bounds by p_k proves (27). □

A.2 WEIGHTED STOCHASTIC GRADIENT METHOD

Write $\mathcal{A} = [-B-1, B]^K$ for the shift component of the feasible set $\mathcal{Z}$ in Section 3. We use the weighted norm, stochastic gradient, projected update, and constants C, V, D defined there. For a sampled pair (k, X), the function differentiated below is

$$R(\theta) + Kp_k\left[s_k - 1 + h_\rho(L_k(X,\theta) - s_k)\right].$$

Lemma A.2. *Every sample function is convex and C/ρ-smooth in the weighted norm. If $\rho\bar{\kappa} \le 1/8$, z_ρ minimizes G_ρ, and $g(z_\rho)$ is a stochastic gradient at that point, then*

$$\mathbb{E}\,\|g(z_\rho)\|^2_{P^{-1}} \le V.$$

Consequently, at every iterate,

$$\mathbb{E}_t\,\|g_t\|^2_{P^{-1}} \le \frac{4C}{\rho}[G_\rho(z_t) - G_\rho(z_\rho)] + 2V. \tag{29}$$

Proof. For a direction (u, v), the second directional derivative of one sample function is

$$\begin{aligned} &u^\top \nabla^2 R(\theta) u + Kp_k w\, u^\top \nabla^2 L_k(X,\theta) u \\ &\qquad + Kp_k h''_\rho(L_k - s_k)[\langle \nabla L_k, u\rangle - v_k]^2. \end{aligned}$$

All terms are nonnegative. Weighted Cauchy–Schwarz gives

$$[\langle \nabla L_k, u\rangle - v_k]^2 \le \left(G^2 + \frac{1}{p_k}\right)(\|u\|^2 + p_k v_k^2).$$

Together with $w \le 1/\rho$ and $h''_\rho \le 1/(4\rho)$, this bounds the directional derivative by $(C/\rho)\,\|(u,v)\|^2_P$.

At z_ρ, let $w_{k,*}$ be the weight in group k. The shift first-order condition gives $\mathbb{E}_k w_{k,*} = 1$. Proposition A.1 gives

$$w_{k,*} \le q_k W_k, \qquad q_k \le (1-\rho\bar{\kappa})^{-1} \le \frac{8}{7},$$

where $W_k = e^{L_k - F_k}$ and $q_k = e^{F_k - s_{k,\rho}}$. Consequently,

$$\mathbb{E}_k w_{k,*}^2 \le q_k^2 \mathbb{E}_k W_k^2 \le \frac{64}{49}\bar{\kappa} < 2\bar{\kappa} \le 4\bar{\kappa} \qquad \text{when } \rho\bar{\kappa} \le 1/8.$$

Substituting these two bounds in the stochastic gradients of Section 3 and averaging over the uniformly sampled group gives the value of V in (4).

We prove (29) directly using the globally smooth scalar softplus. Fix $z = (\theta, s)$ in the feasible set and write $z_* := z_\rho = (\theta_*, s_*)$. Using the same sampled pair (k, X) for the current and optimal weights, set

$$a = L_k(X,\theta) - s_k, \qquad a_* = L_k(X,\theta_*) - s_{k,*}, \qquad w = h_\rho'(a), \qquad w_* = h_\rho'(a_*).$$

For a differentiable function F, write $D_F(x,y) := F(x) - F(y) - \langle \nabla F(y), x - y\rangle$. In particular, with $h = h_\rho$,

$$D_h(a, a_*) = h_\rho(a) - h_\rho(a_*) - w_*(a - a_*).$$

Since h_ρ is convex and $(4\rho)^{-1}$-smooth on all of $\mathbb{R}$, scalar co-coercivity gives

$$(w - w_*)^2 \le \frac{D_h(a, a_*)}{2\rho}.$$

Decompose the sampled Euclidean gradient as $g(z) = b + r$, where

$$b = \big(\nabla R(\theta) + K p_k w_* \nabla L_k(X,\theta),\ K p_k (1 - w_*) e_k\big),$$
$$r = K p_k (w - w_*)\big(\nabla L_k(X,\theta), -e_k\big),$$

and e_k is the kth coordinate vector in $\mathbb{R}^K$. The previously established identities $\mathbb{E}_k w_* = 1$ and $\mathbb{E}_k w_*^2 \le 4\bar{\kappa}$, together with the uniform gradient bounds and $\sum_k p_k^2 \le 1$, imply

$$\begin{aligned}\mathbb{E}\,\|b\|_{P^{-1}}^2 &\le 2G_R^2 + 2KG^2 \sum_k p_k^2 \mathbb{E}_k w_*^2 + K \sum_k p_k \mathbb{E}_k (1 - w_*)^2 \\ &\le 2G_R^2 + 8K\bar{\kappa}G^2 + 4K\bar{\kappa} - K \le V.\end{aligned}$$

Moreover, since $p_k \le 1$,

$$\begin{aligned}\mathbb{E}\,\|r\|_{P^{-1}}^2 &\le K(G^2+1) \sum_k p_k \mathbb{E}_k (w - w_*)^2 \\ &\le \frac{K(G^2+1)}{2\rho} \sum_k p_k \mathbb{E}_k D_h(a, a_*).\end{aligned}$$

The Bregman divergence of the full objective satisfies the exact identity

$$D_{G_\rho}(z, z_*) = D_R(\theta, \theta_*) + \sum_k p_k \mathbb{E}_k \big[D_h(a, a_*) + w_* D_{L_k(X,\cdot)}(\theta, \theta_*)\big].$$

Here the same sampled X is used in both arguments of $D_{L_k(X,\cdot)}$. Convexity of R and every $L_k(X,\cdot)$, and $w_* > 0$, give

$$D_{G_\rho}(z, z_*) \ge \sum_k p_k \mathbb{E}_k D_h(a, a_*).$$

Consequently,

$$\begin{aligned}\mathbb{E}\,\|g(z)\|_{P^{-1}}^2 &\le 2\mathbb{E}\,\|b\|_{P^{-1}}^2 + 2\mathbb{E}\,\|r\|_{P^{-1}}^2 \\ &\le 2V + \frac{K(G^2+1)}{\rho} D_{G_\rho}(z, z_*) \\ &\le 2V + \frac{4C}{\rho}[G_\rho(z) - G_\rho(z_*)].\end{aligned}$$

The last inequality uses $4C \ge K(G^2+1)$ and constrained first-order optimality, $\langle \nabla G_\rho(z_*), z - z_*\rangle \ge 0$. Conditioning on the history through z_t and taking a fresh sample gives (29). □

A.3 CURVATURE OF THE JOINT SURROGATE

The pointwise second derivative of softplus can be very small. The next scalar inequality instead measures curvature relative to the optimal shift.

Lemma A.3. *Fix a_* and let*

$$w_* := w_\rho(a_*), \qquad m_* := w_*[1 - \rho w_*].$$

Then, for every $a \in \mathbb{R}$,

$$[w_\rho(a) - w_*](a - a_*) \geq m_* \frac{(a - a_*)^2}{1 + |a - a_*|}. \tag{30}$$

Proof. Put $d = a - a_*$. If $d \geq 0$, direct substitution gives

$$w_\rho(a_* + d) - w_* = \frac{m_*(e^d - 1)}{1 + \rho w_*(e^d - 1)} \geq m_*(1 - e^{-d}).$$

If $d = -x \leq 0$, the same calculation gives

$$w_* - w_\rho(a_* - x) = \frac{m_*(1 - e^{-x})}{1 - \rho w_*(1 - e^{-x})} \geq m_*(1 - e^{-x}).$$

The weight is increasing, and $1 - e^{-x} \geq x/(1 + x)$ for $x \geq 0$. Multiplying by $|d|$ proves (30). □

At the surrogate minimizer, the average scalar curvature is not small. Indeed, $\mathbb{E}_k w_{k,*} = 1$ and $\mathbb{E}_k w_{k,*}^2 \leq 4\bar{\kappa}$, so

$$\mathbb{E}_k\{w_{k,*}[1 - \rho w_{k,*}]\} = 1 - \rho \mathbb{E}_k w_{k,*}^2 \geq \frac{1}{2}. \tag{31}$$

We now state the only additional assumption needed for the fast generic rate. The profiled objective J_ρ is called γ-strongly convex if

$$J_\rho(\theta') \geq J_\rho(\theta) + \langle \nabla J_\rho(\theta), \theta' - \theta \rangle + \frac{\gamma}{2} \|\theta' - \theta\|^2 .$$

When J_ρ is twice differentiable, a sufficient condition is $\nabla^2 J_\rho(\theta) \succeq \gamma I$ on Θ. The optimal shift is unique and interior, and its Hessian block is positive definite. Differentiating its first-order condition and then the envelope identity shows that the profile Hessian is the Schur complement

$$\nabla^2 J_\rho = G_{\theta\theta} - G_{\theta s} G_{ss}^{-1} G_{s\theta}, \tag{32}$$

where all blocks are evaluated at the optimal shifts. Thus profile curvature concerns the parameter after the auxiliary variables have adjusted optimally.

Lemma A.4. *Suppose $\rho\bar{\kappa} \leq 1/8$ and J_ρ is γ-strongly convex. Define*

$$\mu := \left[\frac{2(1 + 2G^2)}{\gamma} + 16B + 8 \right]^{-1}. \tag{33}$$

Then, for every $z \in \Theta \times \mathcal{A}$,

$$\langle \nabla G_\rho(z) - \nabla G_\rho(z_\rho), z - z_\rho \rangle \geq \mu \|z - z_\rho\|_P^2 . \tag{34}$$

Proof. Write $u = \theta - \theta_\rho$ and $v_k = s_k - s_{k,\rho}$. For an observation in group k, put

$$d_k = L_k(X, \theta) - L_k(X, \theta_\rho) - v_k, \qquad m_k = w_{k,*}(1 - \rho w_{k,*}).$$

Indeed, $|L_k(X, \theta) - L_k(X, \theta_\rho)| \leq 2B$ and $|v_k| \leq 2B + 1$, so $|d_k| \leq 4B + 1$. Convexity of each loss and Lemma A.3 give the needed samplewise lower bound as follows. Write

$$a = L_k(X, \theta) - s_k, \qquad a_* = L_k(X, \theta_\rho) - s_{k,\rho}.$$

The two convexity inequalities for L_k imply

$$\begin{aligned} &\langle w_\rho(a) \nabla L_k(X, \theta) - w_\rho(a_*) \nabla L_k(X, \theta_\rho), u \rangle - [w_\rho(a) - w_\rho(a_*)] v_k \\ &\quad \geq [w_\rho(a) - w_\rho(a_*)][L_k(X, \theta) - L_k(X, \theta_\rho) - v_k] \\ &\quad = [w_\rho(a) - w_\rho(a_*)] d_k. \end{aligned}$$

Lemma A.3 and $|d_k| \le 4B + 1$ bound this by $m_k d_k^2/(4B + 2)$. After taking expectations and weighting the groups, the softplus contribution to the gradient secant is therefore at least

$$\frac{A}{4B + 2}, \qquad A := \sum_{k=1}^{K} p_k \mathbb{E}_k(m_k d_k^2).$$

The regularizer is convex, so its contribution is nonnegative. Hence, if S denotes the left-hand side of (34),

$$S \ge \frac{A}{4B + 2}. \tag{35}$$

Next, $v_k = L_k(X, \theta) - L_k(X, \theta_\rho) - d_k$ and $|L_k(X, \theta) - L_k(X, \theta_\rho)| \le G \|u\|$. Multiply $(r + s)^2 \le 2r^2 + 2s^2$ by m_k and take expectations to obtain

$$v_k^2 \mathbb{E}_k m_k \le 2G^2 \|u\|^2 \mathbb{E}_k m_k + 2\mathbb{E}_k(m_k d_k^2).$$

Equation (31) gives $\mathbb{E}_k m_k \ge 1/2$ (and trivially $\mathbb{E}_k m_k \le 1$), so division by $\mathbb{E}_k m_k$ and averaging with the p_k gives

$$\sum_{k=1}^{K} p_k v_k^2 \le 4A + 2G^2 \|u\|^2,$$

and therefore

$$\|z - z_\rho\|_P^2 \le (1 + 2G^2) \|u\|^2 + 4A. \tag{36}$$

The symmetric Bregman divergence of G_ρ is the sum of its two one-sided Bregman divergences and equals S. Convexity makes both summands nonnegative, so S is at least either one. At the optimal shifts, $\nabla_s G_\rho(z_\rho) = 0$ and $\nabla_\theta G_\rho(z_\rho) = \nabla J_\rho(\theta_\rho)$. Since $G_\rho(\theta, s) \ge J_\rho(\theta)$ and equality holds at z_ρ, the one-sided Bregman divergence satisfies

$$G_\rho(z) - G_\rho(z_\rho) - \langle \nabla G_\rho(z_\rho), z - z_\rho \rangle \ge J_\rho(\theta) - J_\rho(\theta_\rho) - \langle \nabla J_\rho(\theta_\rho), u \rangle .$$

Strong convexity of J_ρ now gives

$$S \ge \frac{\gamma}{2} \|u\|^2 . \tag{37}$$

Equations (35) and (37) imply

$$\|z - z_\rho\|_P^2 \le \left[\frac{2(1 + 2G^2)}{\gamma} + 4(4B + 2) \right] S.$$

This is (34). □

Remark A.5. If R is λ-strongly convex, then $G_\rho(\theta, s) - \lambda \|\theta\|^2 /2$ is jointly convex. Partial minimization over s preserves convexity, so J_ρ remains λ-strongly convex and one may take $\gamma = \lambda$. The Cox application below instead obtains γ from the partial likelihood itself, and therefore permits $\lambda = 0$.

A.4 Generic convergence bounds

Proof of Theorem 3.1. Write $\bar{z}_T = T^{-1} \sum_{t=1}^{T} z_t$. Nonexpansiveness of the weighted projection, unbiasedness, convexity, and Lemma A.2 give

$$\mathbb{E}_t \|z_{t+1} - z_\rho\|_P^2 \le \|z_t - z_\rho\|_P^2 - \left(2\eta_T - \frac{4C\eta_T^2}{\rho_T} \right) [G_\rho(z_t) - G_\rho(z_\rho)] + 2\eta_T^2 V.$$

The chosen step size makes the coefficient in parentheses equal to η_T. Sum over t, use $\|z_1 - z_\rho\|_P^2 \le D^2$, and apply Jensen's inequality to the averaged iterate. This yields

$$\mathbb{E}[G_\rho(\bar{z}_T) - G_\rho(z_\rho)] \le \frac{4CD^2}{\rho_T T} + \frac{\rho_T V}{2C}.$$

Proposition A.1 transfers this inequality from G_ρ to J. Since $\rho_T \bar{\kappa} \le 1/8$, the approximation error is at most $-\rho_T/2 - \log(1 - \rho_T \bar{\kappa}) \le 8\rho_T \bar{\kappa}/7$. □

The fast result is most useful in distance form, because applications can compare the surrogate minimizer with the desired target through a problem-specific gradient bound.

Proof of Theorem 3.2. Let z_ρ minimize G_ρ. As in the convex proof, nonexpansiveness and unbiasedness give

$$\mathbb{E}_t \|z_{t+1} - z_\rho\|_P^2 \le \|z_t - z_\rho\|_P^2 - 2\eta \langle \nabla G_\rho(z_t), z_t - z_\rho \rangle + \eta^2 \mathbb{E}_t \|g_t\|_{P^{-1}}^2 .$$

The gradient inner product is at least the surrogate gap by convexity and at least $\mu \|z_t - z_\rho\|_P^2$ by Lemma A.4 and first-order optimality. It is therefore at least half the sum of these two lower bounds. Substitute this fact and (29). Since $\eta \le \rho/(4C)$, the coefficient of the nonnegative surrogate gap is nonpositive and may be dropped, giving

$$\mathbb{E}_t \|z_{t+1} - z_\rho\|_P^2 \le (1 - \mu\eta) \|z_t - z_\rho\|_P^2 + 2\eta^2 V.$$

Because $0 \le 1 - \mu\eta < 1$, iteration of this recursion and the geometric sum give

$$\mathbb{E} \|z_{T+1} - z_\rho\|_P^2 \le (1 - \mu\eta)^T D^2 + 2\eta^2 V \sum_{j=0}^{T-1} (1 - \mu\eta)^j \le e^{-\mu\eta T} D^2 + \frac{2\eta V}{\mu},$$

which proves (8). For the displayed schedule, $\eta_T = \rho_T/(4C)$, $\mu\eta_T T = 2 \log T$, and $\eta_T \le 1/\mu$ for $T \ge 2$. Substitution proves (9) after dropping the shift coordinates. This last step is the standard constant-step strongly convex SGD recursion (Moulines & Bach, 2011). □

Proof of Corollary 3.3. Let $\theta^\star$ minimize J. By the value approximation (27), optimality and the two sides of that bound give

$$J_\rho(\theta^\star) - J_\rho(\theta_\rho) \le J(\theta^\star) - J(\theta_\rho) - \rho/2 - \log(1 - \rho\bar{\kappa}) \le -\rho/2 - \log(1 - \rho\bar{\kappa}).$$

Strong convexity of J_ρ therefore gives

$$\|\theta_\rho - \theta^\star\|^2 \le \frac{2}{\gamma}[-\rho/2 - \log(1 - \rho\bar{\kappa})] \le \frac{16\rho\bar{\kappa}}{7\gamma}.$$

The smoothness assumptions at the start of this section imply that J is L_J-smooth. Since $\theta^\star$ is stationary, smoothness, the preceding display, (9), and $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ give (10). Finally, $\rho_T = 8C \log T/(\mu T)$ yields the stated order. □

Stationarity is needed because the smoothness upper bound contains a linear term at a constrained boundary. In Cox regression we use a sharper gradient comparison, its softplus contribution is quadratic after conversion to objective error.

B Cox regression

This section proves the uniform comparison in Theorem 4.1 and specializes the generic recursion to obtain Corollary 4.2, with all constants retained.

B.1 Notation and data constants

We use the data, grouping rule, and objectives of Section 4: the exact $\mathcal{L}$, grouped $\widetilde{\mathcal{L}}_\delta$, joint $G_{\rho,\delta}$, and profiled $\mathcal{L}_{\rho,\delta}$. The optional ridge coefficient $\lambda \ge 0$ is the same in all four objectives. Throughout this section, $\|x_j\| \le X$ and $|\beta^\top x_j| \le M$ for every subject j and $\beta \in \mathcal{B}$.

For $i \in \mathcal{E}$, let Q_i^β be the Cox distribution on $\mathcal{R}_i$:

$$Q_i^\beta(j) := \frac{e^{\beta^\top x_j}}{\sum_{\ell \in \mathcal{R}_i} e^{\beta^\top x_\ell}}.$$

The constants κ and L_q are defined in (18) and (19). Both are independent of the grouping.

The suprema over $\mathcal{B}$ are needed for a global optimization theorem. For a local statistical comparison they may instead be taken over a fixed neighborhood of the target.

B.2 GROUPING BOUNDS

Recall $r_\delta = \delta/(1+\delta)$ from Section 4. By the grouping rule, at most this fraction of the largest risk set disappears inside one group.

Lemma B.1. *Suppose*

$$r_\delta \le 1 - q, \qquad L_q r_\delta \le \frac{1}{2}. \tag{38}$$

For any two risk sets $\mathcal{R}_v \subset \mathcal{R}_u$ in the same group,

$$|a_u(\beta) - a_v(\beta)| \le 2L_q r_\delta, \tag{39}$$

$$\|\nabla a_u(\beta) - \nabla a_v(\beta)\| \le 2XL_q r_\delta, \tag{40}$$

$$\left\|\nabla^2 a_u(\beta) - \nabla^2 a_v(\beta)\right\|_{\mathrm{op}} \le 5X^2 L_q r_\delta. \tag{41}$$

Proof. Let $A = \mathcal{R}_u \setminus \mathcal{R}_v$ and define its cardinality and Cox-mass fractions by

$$r := \frac{|A|}{|\mathcal{R}_u|}, \qquad \alpha := Q_u^\beta(A).$$

The first condition in (38) makes the pair admissible in (19). Therefore

$$r \le r_\delta, \qquad \alpha \le L_q r \le L_q r_\delta. \tag{42}$$

Removing A changes the average risk weight according to the exact identity

$$a_u - a_v = \log \frac{1-r}{1-\alpha}. \tag{43}$$

Both r and α lie in $[0, L_q r_\delta]$. Hence

$$|a_u - a_v| \le -\log(1 - L_q r_\delta) \le 2L_q r_\delta,$$

which proves (39).

Let Q_A be the Cox distribution conditional on A. The outer Cox law is the mixture

$$Q_u^\beta = (1-\alpha)Q_v^\beta + \alpha Q_A.$$

Its mean therefore differs from that of Q_v^β by at most $2X\alpha$, which proves (40).

Write H_u, H_v, H_A for the corresponding covariance matrices and g_v, g_A for the two conditional means. The covariance of a mixture is

$$H_u = (1-\alpha)H_v + \alpha H_A + \alpha(1-\alpha)(g_A - g_v)(g_A - g_v)^\top.$$

All covariance matrices are between 0 and $X^2 I$, while $\|g_A - g_v\| \le 2X$. Consequently

$$\|H_u - H_v\|_{\mathrm{op}} \le \alpha X^2 + 4\alpha X^2 \le 5X^2 L_q r_\delta.$$

This proves the last claim. □

Proposition B.2. *Under (38), for every $\beta \in \mathcal{B}$,*

$$\left\|\nabla \widetilde{\mathcal{L}}_\delta(\beta) - \nabla \mathcal{L}(\beta)\right\| \le XL_q^2 r_\delta^2, \tag{44}$$

$$\nabla^2 \widetilde{\mathcal{L}}_\delta(\beta) \succeq \nabla^2 \mathcal{L}(\beta) - \frac{5}{2} X^2 L_q^2 r_\delta^2 I. \tag{45}$$

Proof. Fix a group and suppress β in $a_i(\beta)$. Write $g_i = \nabla a_i(\beta)$ and $H_i = \nabla^2 a_i(\beta)$. Differentiating (13) gives

$$\nabla \Phi_k = \sum_{i \in \mathcal{I}_k} \pi_i g_i, \tag{46}$$

$$\nabla^2 \Phi_k = \sum_{i \in \mathcal{I}_k} \pi_i H_i + \mathrm{Cov}_\pi(g_i), \tag{47}$$

where $\pi_i = e^{a_i}/\sum_{r\in\mathcal{I}_k} e^{a_r}$. The exact Cox group uses the uniform average of the g_i and H_i.

By Lemma B.1, the oscillation of the a_i is at most $2L_q r_\delta$, the diameter of the g_i is at most $2XL_q r_\delta$, and the diameter of the H_i is at most $5X^2 L_q r_\delta$. To compare the group weights, interpolate from the uniform law u to π via $\pi_{t,i} = e^{ta_i}/\sum_{r\in\mathcal{I}_k} e^{ta_r}$, $0 \le t \le 1$. For any scalar array f_i, write $\operatorname{osc}(f) = \max_i f_i - \min_i f_i$. Then $\frac{d}{dt}\mathbb{E}_{\pi_t} f = \operatorname{Cov}_{\pi_t}(f, a)$. Cauchy–Schwarz and the variance bound $\operatorname{Var}(f) \le \operatorname{osc}(f)^2/4$ therefore give

$$|\mathbb{E}_\pi f - \mathbb{E}_u f| \le \frac{1}{4}\operatorname{osc}(f)\operatorname{osc}(a). \tag{48}$$

Apply this to $f_i = v^\top g_i$ and take the supremum over unit vectors v. The groupwise gradient difference is at most

$$\left\| \sum_i \pi_i g_i - \frac{1}{m_k}\sum_i g_i \right\| \le (2XL_q r_\delta)\frac{2L_q r_\delta}{4} = XL_q^2 r_\delta^2.$$

For the Hessian, apply (48) to $f_i = v^\top H_i v$ for each unit v. The covariance term in (47) is positive semidefinite, so

$$\nabla^2\Phi_k \succeq \frac{1}{m_k}\sum_i H_i - (5X^2 L_q r_\delta)\frac{2L_q r_\delta}{4} I.$$

Weighting the groupwise bounds by p_k proves the proposition. □

The same oscillation bound removes the group dependence from the softplus moment.

Lemma B.3. *Under (38), every group satisfies*

$$\sup_{\beta\in\mathcal{B}} \mathbb{E}_k \exp\big(2[\beta^\top x_J - \Phi_k(\beta)]\big) \le \frac{9}{8}\kappa. \tag{49}$$

Proof. Put

$$A_i = e^{a_i(\beta)}, \qquad B_i = \frac{1}{n_i}\sum_{j\in\mathcal{R}_i} e^{2\beta^\top x_j}.$$

Definition (18) gives $B_i \le \kappa A_i^2$. Consequently the group moment is at most

$$\kappa\frac{m_k^{-1}\sum_i A_i^2}{(m_k^{-1}\sum_i A_i)^2}.$$

If positive numbers y lie in $[a, b]$, then $(y-a)(b-y) \ge 0$ implies $y^2 \le (a+b)y - ab$. Writing $\overline{y} = \mathbb{E}y$ and maximizing over $\overline{y} \in [a, b]$ gives

$$\frac{\mathbb{E}y^2}{(\mathbb{E}y)^2} \le \frac{a+b}{\overline{y}} - \frac{ab}{\overline{y}^2} \le \frac{(a+b)^2}{4ab}$$

the last maximum occurs at $\overline{y} = 2ab/(a+b)$. For any pair in the group, identity (43) shows that both $e^{a_u - a_v}$ and $e^{a_v - a_u}$ are at most $(1 - L_q r_\delta)^{-1}$. Hence the largest and smallest group values obey the sharper ratio bound $b/a = e^{\operatorname{osc} a} \le (1 - L_q r_\delta)^{-1} \le 2$. The last display is therefore at most $(1+2)^2/(4\cdot 2) = 9/8$. □

B.3 Softplus gradient and curvature

The next lemma is stated for a single group distribution. It will be applied with the fixed envelope from Lemma B.3. Its $\bar\kappa$ is the group-moment envelope of Section 3, bounded here by $9\kappa/8$.

Lemma B.4. *Let*

$$\Phi(\beta) = \log \mathbb{E}e^{\beta^\top X}$$

for a distribution supported on $\|X\| \le X_0$, and suppose

$$\mathbb{E}e^{2[\beta^\top X - \Phi(\beta)]} \le \bar\kappa.$$

Let Φ_ρ be its profiled softplus representation. If $\rho\bar\kappa \le 1/8$, then

$$\|\nabla\Phi_\rho(\beta) - \nabla\Phi(\beta)\| \le 3X_0\rho\bar\kappa, \tag{50}$$

$$\nabla^2\Phi_\rho(\beta) \succeq \nabla^2\Phi(\beta) - 16X_0^2\rho\bar\kappa I. \tag{51}$$

Proof. Put $W = e^{\beta^\top X - \Phi(\beta)}$, so $\mathbb{E}W = 1$ and $\mathbb{E}W^2 \le \bar{\kappa}$. At the optimal shift, write

$$c = e^{\Phi(\beta) - s_\rho(\beta)}, \qquad w = \frac{cW}{1 + \rho cW}.$$

The shift first-order condition is $\mathbb{E}w = 1$, and Proposition A.1 gives

$$1 < c \le \frac{1}{1 - \rho\bar{\kappa}}.$$

Therefore

$$\begin{aligned}\mathbb{E}|w - W| &\le (c-1)\mathbb{E}W + \rho c\mathbb{E}W^2 \\ &\le \frac{2\rho\bar{\kappa}}{1 - \rho\bar{\kappa}} \le 3\rho\bar{\kappa}.\end{aligned}$$

Since the two gradients are $\mathbb{E}(wX)$ and $\mathbb{E}(WX)$, this proves (50).

For the Hessian, let

$$m = w(1 - \rho w).$$

The exact and profiled Hessians have the weighted least-squares forms

$$v^\top \nabla^2 \Phi(\beta) v = \min_{r \in \mathbb{R}} \mathbb{E}[W(v^\top X - r)^2], \tag{52}$$

$$v^\top \nabla^2 \Phi_\rho(\beta) v = \min_{r \in \mathbb{R}} \mathbb{E}[m(v^\top X - r)^2]. \tag{53}$$

For the first identity, $\mathbb{E}W = 1$ and $\nabla^2\Phi = \mathbb{E}(WXX^\top) - \mathbb{E}(WX)\mathbb{E}(WX)^\top$. For the second, the joint softplus Hessian blocks are

$$G_{\beta\beta} = \mathbb{E}(mXX^\top), \qquad G_{\beta s} = -\mathbb{E}(mX), \qquad G_{ss} = \mathbb{E}m.$$

Taking their Schur complement gives $\mathbb{E}(mXX^\top) - \mathbb{E}(mX)\mathbb{E}(mX)^\top / \mathbb{E}m$, which is exactly the minimum over r in (53).

Moreover,

$$\mathbb{E}w^2 \le c^2\mathbb{E}W^2 \le \frac{\bar{\kappa}}{(1 - \rho\bar{\kappa})^2},$$

and hence, with $a = \rho\bar{\kappa} \le 1/8$,

$$\mathbb{E}|m - W| \le \mathbb{E}|w - W| + \rho\mathbb{E}w^2 \le \frac{2a}{1-a} + \frac{a}{(1-a)^2} = \frac{a(3-2a)}{(1-a)^2} \le 4a = 4\rho\bar{\kappa}.$$

Let r_m minimize (53). It is a weighted mean of $v^\top X$ and therefore lies between the smallest and largest values of $v^\top X$. Thus $|v^\top X - r_m| \le 2X_0 \|v\|$. Evaluating both weighted variances at r_m gives

$$\begin{aligned}v^\top \nabla^2 \Phi_\rho v &= \mathbb{E}[m(v^\top X - r_m)^2] \\ &\ge \min_r \mathbb{E}[W(v^\top X - r)^2] - 4X_0^2 \|v\|^2 \mathbb{E}|m - W| \\ &\ge v^\top \nabla^2 \Phi\, v - 16 X_0^2 \rho\bar{\kappa} \|v\|^2.\end{aligned}$$

This proves (51). □

Combining the preceding results gives one finite-sample comparison that will be used twice: first for computation and later for statistics.

Proposition B.5. *Suppose* (38) *holds and*

$$\rho\kappa \le \frac{1}{9}. \tag{54}$$

Then, uniformly over $\beta \in \mathcal{B}$,

$$\|\nabla\mathcal{L}_{\rho,\delta}(\beta) - \nabla\mathcal{L}(\beta)\| \le XL_q^2 r_\delta^2 + 4X\rho\kappa, \tag{55}$$

$$\nabla^2\mathcal{L}_{\rho,\delta}(\beta) \succeq \nabla^2\mathcal{L}(\beta) - \left(\frac{5}{2}X^2L_q^2r_\delta^2 + 20X^2\rho\kappa\right) I. \tag{56}$$

Proof. Lemma B.3 bounds each group moment by $9\kappa/8$. Condition (54) therefore makes $\rho(9\kappa/8) \le 1/8$ in every group. Lemma B.4, followed by averaging with the p_k, gives a softplus gradient error smaller than $4X\rho\kappa$ and a Hessian loss smaller than $20X^2\rho\kappa$. Add the grouping bounds in Proposition B.2. □

B.4 Curvature transfer and deterministic minimizer comparison

We now impose the main Cox assumption:

$$\nabla^2 \mathcal{L}(\beta) \succeq \nu I \qquad \text{for every } \beta \in \mathcal{B}, \tag{57}$$

for some $\nu > 0$. When $\lambda = 0$, this is an identifiability and design condition on the exact partial likelihood, not curvature supplied by a penalty.

Theorem B.6. *Assume* (57), (38), *and* (54). *Let* $\beta^\star$ *minimize* $\mathcal{L}$ *and let* $\beta_{\rho,\delta}$ *minimize* $\mathcal{L}_{\rho,\delta}$ *on* $\mathcal{B}$. *Then*

$$\|\beta_{\rho,\delta} - \beta^\star\| \le \frac{X}{\nu}\left(L_q^2 r_\delta^2 + 4\rho\kappa\right). \tag{58}$$

If, in addition,

$$\frac{5}{2}X^2 L_q^2 r_\delta^2 + 20X^2\rho\kappa \le \frac{\nu}{2}, \tag{59}$$

then

$$\nabla^2 \mathcal{L}_{\rho,\delta}(\beta) \succeq \frac{\nu}{2} I \qquad (\beta \in \mathcal{B}). \tag{60}$$

Proof. Under the additional curvature budget, the curvature conclusion follows immediately from (56). For the distance bound, which does not use that budget, put $h = \beta_{\rho,\delta} - \beta^\star$. Strong convexity of the exact objective gives

$$\langle \nabla\mathcal{L}(\beta_{\rho,\delta}) - \nabla\mathcal{L}(\beta^\star), h\rangle \ge \nu \|h\|^2.$$

The variational inequalities for the two constrained minimizers give

$$\langle \nabla\mathcal{L}(\beta^\star), h\rangle \ge 0, \qquad \langle \nabla\mathcal{L}_{\rho,\delta}(\beta_{\rho,\delta}), h\rangle \le 0.$$

Insert and subtract $\nabla\mathcal{L}_{\rho,\delta}(\beta_{\rho,\delta})$ in the left side of the strong-convexity display. The two variational-inequality terms are nonpositive after this rearrangement, so

$$\nu\|h\|^2 \le \langle \nabla\mathcal{L}(\beta_{\rho,\delta}) - \nabla\mathcal{L}_{\rho,\delta}(\beta_{\rho,\delta}), h\rangle \le \|\nabla\mathcal{L}(\beta_{\rho,\delta}) - \nabla\mathcal{L}_{\rho,\delta}(\beta_{\rho,\delta})\|\,\|h\|.$$

Now use (55). If $h = 0$ there is nothing to prove. □

B.5 End-to-end computational rate

For a declared horizon T, choose the grouping tolerance δ_T, build the groups once, and keep them fixed throughout the run. Let K_T be their number and define

$$A_{K_T} := \lambda + \frac{K_T}{4}(X^2 + 1), \tag{61}$$

$$V_T := 2\lambda^2 B_0^2 + \frac{9}{2}K_T\kappa(5X^2 + 1), \tag{62}$$

$$D^2 := \mathrm{diam}(\mathcal{B})^2 + (2M + 1)^2, \tag{63}$$

where $B_0 := \sup_{\beta\in\mathcal{B}}\|\beta\|$. These are the generic constants (4)–(5), using the fixed group moment envelope $9\kappa/8$.

Under the curvature budget (59), the profiled Cox surrogate has curvature at least $\nu/2$. Taking $\gamma = \nu/2$ in (7) gives the T-independent restricted-secant constant μ stated in Corollary 4.2. This constant has only polynomial dependence on M. The separate intrinsic factors L_q and κ may still be large when risk weights are highly heterogeneous.

The specialized stochastic step samples a group uniformly, then an event uniformly from that group, and finally a subject uniformly from the event's risk set. Its expected objective has

$$R(\beta) = \frac{\lambda}{2}\|\beta\|^2 - \beta^\top \bar{x}_{\mathcal{E}}, \qquad L(\beta, x) = \beta^\top x.$$

The weighted projection and shift box $[-M - 1, M]^{K_T}$ are as in (3). All expectations in this subsection are conditional on the data and are over the random draws made by the algorithm.

Lemma B.7. *Suppose (38) holds and $\rho\kappa \le 1/9$. For the single-triple gradient g_t in (16)–(17), the minimizer z_ρ of $G_{\rho,\delta}$ on $\mathcal{B} \times [-M-1, M]^{K_T}$ satisfies*

$$\mathbb{E}_t \|g_t\|_{P^{-1}}^2 \le \frac{4A_{K_T}}{\rho}[G_{\rho,\delta}(z_t) - G_{\rho,\delta}(z_\rho)] + 2V_T. \tag{64}$$

Proof. Write $\bar{\kappa}_{\mathrm{grp}}$ for the maximum normalized exponential moment of the group distributions. Lemma B.3 gives $\bar{\kappa}_{\mathrm{grp}} \le 9\kappa/8$, so $\rho\bar{\kappa}_{\mathrm{grp}} \le 1/8$. At z_ρ, the groupwise shift condition and the moment bound used in Lemma A.2 give $\mathbb{E}_k w_{k,*} = 1$ and $\mathbb{E}_k w_{k,*}^2 \le 4\bar{\kappa}_{\mathrm{grp}}$. For the same sampled triple at $z = (\beta, s)$ and z_ρ, decompose its gradient as $g(z) = b + r$, where

$$b_\beta = \lambda\beta + K_T p_k(w_{k,*}x_J - x_I), \qquad b_s = K_T p_k(1 - w_{k,*})e_k,$$
$$r = K_T p_k(w - w_{k,*})(x_J, -e_k).$$

The bound $\|\beta\| \le B_0$, $\|x_j\| \le X$, and uniform group sampling give

$$\begin{aligned}\mathbb{E}\|b\|_{P^{-1}}^2 &\le 2\lambda^2 B_0^2 + 4K_T X^2 \sum_k p_k^2(\mathbb{E}_k w_{k,*}^2 + 1) + K_T \sum_k p_k \mathbb{E}_k(1 - w_{k,*})^2 \\ &\le 2\lambda^2 B_0^2 + 4K_T X^2(4\bar{\kappa}_{\mathrm{grp}} + 1) + 4K_T\bar{\kappa}_{\mathrm{grp}} \le V_T.\end{aligned}$$

The last step uses $1 \le \bar{\kappa}_{\mathrm{grp}} \le 9\kappa/8$. As in the proof of Lemma A.2, scalar co-coercivity of h_ρ yields

$$\mathbb{E}\|r\|_{P^{-1}}^2 \le \frac{K_T(X^2+1)}{2\rho} \sum_k p_k \mathbb{E}_k D_{h_\rho}(a, a_*).$$

The sampled event term is affine in β, so its Bregman divergence vanishes. Consequently the sum on the right is at most $D_{G_{\rho,\delta}}(z, z_\rho)$, which is at most the objective gap by constrained optimality. Apply $\|b + r\|_{P^{-1}}^2 \le 2\|b\|_{P^{-1}}^2 + 2\|r\|_{P^{-1}}^2$ and $K_T(X^2+1) \le 4A_{K_T}$ to obtain (64). □

Theorem B.8. *Assume the standing covariate and predictor bounds and (57). Fix $T \ge 2$ and $q \in (0,1)$, and suppose the chosen δ_T satisfies*

$$r_T := \frac{\delta_T}{1 + \delta_T}, \qquad r_T \le 1 - q, \qquad L_q r_T \le \frac{1}{2}.$$

Set

$$\rho_T := \frac{8A_{K_T}\log T}{\mu T}, \qquad \eta_T := \frac{2\log T}{\mu T}. \tag{65}$$

Suppose $\rho_T \le 1$, $\rho_T\kappa \le 1/9$, and

$$\frac{5}{2}X^2 L_q^2 r_T^2 + 20X^2\rho_T\kappa \le \frac{\nu}{2}. \tag{66}$$

Run weighted projected softplus SGD for T iterations and let $\widehat{\beta}_T = \beta_{T+1}$. Then

$$\begin{aligned}\mathbb{E}\left\|\widehat{\beta}_T - \beta^\star\right\|^2 \le \frac{2D^2}{T^2} + \frac{8V_T\log T}{\mu^2 T} \\ + \frac{2X^2}{\nu^2}\left(L_q^2 r_T^2 + 4\rho_T\kappa\right)^2.\end{aligned} \tag{67}$$

If $\beta^\star$ is in the interior of $\mathcal{B}$, then also

$$\mathbb{E}[\mathcal{L}(\widehat{\beta}_T) - \mathcal{L}(\beta^\star)] \le (\lambda + X^2)\left\{\frac{D^2}{T^2} + \frac{4V_T\log T}{\mu^2 T} + \frac{X^2}{\nu^2}\left(L_q^2 r_T^2 + 4\rho_T\kappa\right)^2\right\}. \tag{68}$$

Proof. Proposition B.5 and (66) show that, at the chosen ρ_T, the profiled objective has Hessian at least $\nu I/2$. Its group moment is at most $9\kappa/8$. Lemma A.4 therefore gives the joint restricted-secant constant μ. The sampled-event estimate (64), unbiasedness, and the weighted projection give the same recursion as in the proof of Theorem 3.2, with $C = A_{K_T}$ and $V = V_T$. Thus, for the profiled-surrogate minimizer β_{ρ_T,δ_T},

$$\mathbb{E}\left\|\widehat{\beta}_T - \beta_{\rho_T,\delta_T}\right\|^2 \le \frac{D^2}{T^2} + \frac{4V_T\log T}{\mu^2 T}.$$

Theorem B.6 bounds the squared distance from β_{ρ_T,δ_T} to $\beta^\star$. Apply $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ to obtain (67).

The exact objective is $(\lambda + X^2)$-smooth: each $\nabla^2 a_i$ is a covariance matrix bounded by $X^2 I$. If $\beta^\star$ is interior, then $\nabla\mathcal{L}(\beta^\star) = 0$, and smoothness gives

$$\mathcal{L}(\beta) - \mathcal{L}(\beta^\star) \le \frac{\lambda + X^2}{2} \|\beta - \beta^\star\|^2 .$$

Combining this with (67) proves (68). □

We now make the dependence on T explicit. Suppose every event risk set has at least cN subjects for a fixed $c \in (0,1)$. Maximal grouping gives

$$K_T \le \min\left\{m, \left\lceil \frac{\log(1/c)}{\log(1+\delta_T)} \right\rceil \right\}. \tag{69}$$

Indeed, let n_k^{start} be the first risk-set size in group k. Maximality gives $n_k^{\text{start}}/n_{k+1}^{\text{start}} > 1+\delta_T$ for $k = 1, \dots, K_T - 1$. Since $n_1^{\text{start}}/n_{K_T}^{\text{start}} \le 1/c$, multiplication yields $(1+\delta_T)^{K_T-1} < 1/c$, which is equivalent to (69) after taking integer parts.

Corollary B.9. *Suppose every event risk set has at least cN subjects for a fixed $c \in (0,1)$, and suppose the constants in Theorem B.8, together with c, do not depend on T. For all sufficiently large T, choose*

$$\delta_T := \left(\frac{\log T}{T}\right)^{1/5} \tag{70}$$

and use (65). Then

$$\mathbb{E}\left\|\widehat{\beta}_T - \beta^\star\right\|^2 = O\left(\left(\frac{\log T}{T}\right)^{4/5}\right), \tag{71}$$

$$\mathbb{E}[\mathcal{L}(\widehat{\beta}_T) - \mathcal{L}(\beta^\star)] = O\left(\left(\frac{\log T}{T}\right)^{4/5}\right) \tag{72}$$

when $\beta^\star$ is interior. Moreover,

$$\mathbb{E}\left\|\widehat{\beta}_T - \beta^\star\right\| = O\left(\left(\frac{\log T}{T}\right)^{2/5}\right). \tag{73}$$

Proof. For $\delta_T \le 1$, (69) gives $K_T = O(\delta_T^{-1})$. Hence

$$A_{K_T} = O(\delta_T^{-1}), \qquad V_T = O(\delta_T^{-1}), \qquad \rho_T = O\left(\frac{\log T}{\delta_T T}\right).$$

With (70), ρ_T is $O((\log T/T)^{4/5})$, so all the smallness and curvature conditions hold eventually. The stochastic term in (67) is

$$O\left(\frac{\log T}{\delta_T T}\right) = O\left(\left(\frac{\log T}{T}\right)^{4/5}\right).$$

The grouping part of the squared minimizer bias is $O(\delta_T^4)$ and has the same order. The squared softplus bias is $O(\rho_T^2)$ and is smaller. This proves (71). The objective result follows from Theorem B.8. Jensen's inequality gives (73). □

Remark B.10. For a fixed finite data set, decreasing δ_T eventually produces singleton event groups. At that point $K_T \le m$ and the grouping bias is zero, so the bound crosses over from the grouping-limited $T^{-4/5}$ regime toward the ordinary $\widetilde{O}(m/T)$ finite-problem rate.

Remark B.11. Suppose instead that $\beta^\star$ is unique and interior and only

$$\nabla^2\mathcal{L}(\beta^\star) \succeq \nu_* I$$

is known. Continuity gives a neighborhood on which $\mathcal{L}(\beta) - \mathcal{L}(\beta^\star) \ge (\nu_*/4)\|\beta - \beta^\star\|^2$. On the compact complement of that neighborhood, uniqueness gives a strictly positive minimum objective gap. Combining the two regions shows that convergence in objective implies convergence in argument, and an expected objective rate also gives the same rate for expected squared distance. This local observation does not by itself establish the uniform profile curvature needed by the RSI during the whole optimization path. That is why Theorem B.8 uses (57) on $\mathcal{B}$.

C STATISTICAL CONSEQUENCE

This section proves Theorem 4.3. To clarify its minimum-risk-set condition, let T^0 and C be a subject's event and censoring times, and put $\widetilde{T} = \min(T^0, C)$, $\Delta = \mathbf{1}\{T^0 \le C\}$, and $Y(t) = \mathbf{1}\{\widetilde{T} \ge t\}$. Only observed failures at or before a fixed horizon $t_{\max} < \infty$ enter the empirical partial likelihood. If $\inf_{0 \le t \le t_{\max}} \Pr\{Y(t) = 1\} > 0$, a uniform law of large numbers for independent subjects makes every risk set through this horizon contain a fixed positive fraction of the sample with probability tending to one. Independent subjects, the proportional-hazards model, conditionally independent censoring, bounded covariates, and nonsingular population information are familiar sufficient conditions for the full estimator's classical limit (Tsiatis, 1981; Andersen & Gill, 1982). As in Section 4.4, we take that limit as given.

Proof of Theorem 4.3. Condition (25) and $L_{q,N} \ge 1$ imply $\delta_N \to_p 0$, $L_{q,N}\delta_N \to_p 0$, and $\rho_N \kappa_N \to_p 0$. Since q is fixed, the smallness conditions of Theorem 4.1 therefore hold with probability tending to one. On the same event, the segment joining the two estimators lies in the convex neighborhood U. Integrating the Hessian bound on that segment and using the variational inequalities for the two constrained minimizers, as in Theorem B.6, gives

$$\left\| \widehat{\beta}_N^{\rho,\delta} - \widehat{\beta}_N \right\| \le \frac{X}{\nu_0} \left(L_{q,N}^2 r_{\delta_N}^2 + 4\rho_N \kappa_N \right) \le \frac{X}{\nu_0} \left(L_{q,N}^2 \delta_N^2 + 4\rho_N \kappa_N \right).$$

After multiplication by $\sqrt{N}$, this tends to zero in probability. Slutsky's theorem transfers the assumed limit of $\widehat{\beta}_N$ to $\widehat{\beta}_N^{\rho,\delta}$.

For the schedule in Theorem 4.3, the two terms of (25) are respectively $O_p(\ell_N^{-2})$ and $O_p(\ell_N^{-1})$. If every event risk set has at least cN subjects with probability tending to one, then (69) and $\log(1+\delta_N) \ge \delta_N/2$ eventually give $K_N = O_p(\delta_N^{-1}) = O_p(N^{1/4}\ell_N) = o_p(N)$ when $\ell_N = o(N^{3/4})$. □

Vanishing ridge. A fixed ridge coefficient changes the population target. If the exact and grouped objectives share a coefficient λ_N and the ridge and unregularized estimators lie in U, the same curvature argument gives $\left\| \widehat{\beta}_{N,\lambda_N} - \widehat{\beta}_{N,0} \right\| \le B_0 \lambda_N / \nu_0$. Thus $\sqrt{N}\lambda_N \to 0$ preserves the ordinary Cox limit.

Numerical accuracy and the two limits. Here N is sample size, while T is the iteration horizon. The schedule in Theorem 4.3 controls statistical approximation as N grows. Corollary B.9 balances approximation and computation as T grows. Once the statistical condition holds, a numerical iterate $\widehat{\beta}_{N,T_N}$ retains the exact Cox limit if $\left\| \widehat{\beta}_{N,T_N} - \widehat{\beta}_N^{\rho,\delta} \right\| = o_p(N^{-1/2})$. A sufficient unconditional criterion is $N\mathbb{E}\left\| \widehat{\beta}_{N,T_N} - \widehat{\beta}_N^{\rho,\delta} \right\|^2 \to 0$. When the grouped estimator and algorithm use the same balanced δ and ρ, the global computational assumptions hold with probability tending to one, and the conditional error prefactor is $O_p(1)$, condition (26) suffices in probability. An unconditional mean-square conclusion additionally requires integrable bounds, for example deterministic computational assumptions and a uniformly integrable prefactor.

D EXPERIMENTAL PROTOCOL

D.1 DATA PREPARATION

For SUPPORT2 (Vanderbilt Biostatistics, n.d.), we use time to death and 14 covariates: age, sex, race, number of comorbidities, diabetes, dementia, cancer status, mean blood pressure, heart rate, respiratory rate, temperature, white blood cell count, sodium, and creatinine. Missing race is a category, records missing another selected covariate or the outcome are excluded, retaining 8,873 of 9,105 subjects. Prognostic scores and outcome-derived variables are excluded. For NWTCO (R survival package documentation, n.d.), we use time to relapse and the corresponding event indicator. Predictors are histology assessments from the local institution and central laboratory, stage, study, and age. Identifiers and the subcohort indicator are excluded.

Synthetic event times follow

$$T = \{E \exp(-x^\top \beta_*)\}^{1/1.5}, \qquad E \sim \mathrm{Exp}(1), \quad x \sim \mathcal{N}(0, \Sigma).$$

Independently, we draw $b \sim \mathcal{N}(0, I_d)$ and set $\beta_{\mathrm{white}} = 0.5b/\|b\|_2$. The correlated cases use $\Sigma_{uv} = r^{|u-v|}$, with $r = 0.95$ for $(N, d) = (10^5, 10)$ and $r = 0.90$ for $(10^6, 20)$. Let $Z \in \mathbb{R}^{N \times d}$ have independent standard normal entries. We form $X = ZL^\top$, $LL^\top = \Sigma$, and set $\beta_* = L^{-\top}\beta_{\mathrm{white}}$. Independent 100k uses $(N, d) = (10^5, 20)$, $\Sigma = I_{20}$, and $\beta_* = \beta_{\mathrm{white}}$. Thus the population standard deviation of the linear predictor is 0.5 in every case. Independent exponential censoring targets 90% in the correlated cases and 30% in Independent 100k, calibrated on separate 65,536-observation pilots; realized fractions are 89.991%, 89.9774%, and 30.131%. Data seeds are 1701 for both correlated cases and 2701 for the independent case. The correlated cases differ in dimension and correlation, as well as N.

All methods share event-stratified 60/20/20 splits, using seed 1701 for training and 1702 for splitting the remainder. Imputation medians, categorical vocabularies, and standardization are fitted on training data. After one-hot encoding, columns with training standard deviation at most 10^{-12} are removed. Validation and test use the fitted transformation unchanged. Sorting and risk indices are shared across all fits within each dataset and training phase. Breslow ties share risk sets while retaining event multiplicities in both the objective and sampling distribution. Table 1 gives training event counts and the resulting numbers of groups at $\delta = 0.05$, alongside dimensions and budgets.

Each dataset uses one realization and one fixed split. Dataset and candidate-menu choices were informed by exploratory optimization results.

D.2 Native objectives and optimizer conventions

All methods use $x^\top\beta$, initialize $\beta = 0$, and add the explicit ridge penalty $\lambda\|\beta\|_2^2/2$, with $\lambda = 10^{-3}$ and zero optimizer weight decay. Their sampled risk sets and native training objectives differ. Every trajectory is evaluated using the same regularized full-Cox objective defined below.

Our method. Groups are sampled uniformly, each with probability $1/K$. We use mini-batches of independent stochastic gradient samples, the unbiased sampled event mean Kp_kx_i with $i \sim \mathrm{Unif}(\mathcal{I}_k)$, weighted projection and arithmetic averaging of the projected post-update coefficient iterates. We project the coefficient vector onto the Euclidean ball $\{\beta : \|\beta\|_2 \le B_0\}$ and clip the updated auxiliary shifts to $[-S, S]$. We fix $\rho = 10^{-4}$, $\delta = 0.05$, $B_0 = 10$, $S = 40$, and $t_0 = 1000$. Batch size and initial step are tuned. Full-Cox diagnostics never enter updates.

Batch LSE. Adam averages over Q sampled events, each with c controls from its full training risk set, the loss

$$-z_i + \log\left(\frac{1}{c}\sum_{r=1}^{c} e^{z_{j_r}}\right), \qquad z_j = x_j^\top\beta,$$

plus ridge. Taking the logarithm of a sampled mean generally gives a biased full-Cox gradient.

Minibatch Cox. Adam uses event-normalized Cox loss with risk sets restricted to shuffled observation batches of size b, retaining incomplete final batches. A batch with no observed events has zero data gradient; the ridge term and the optimizer's existing moment state remain active.

BigSurv. Updates average event-sum losses over ℓ strata of 20 observations (Tarkhan & Simon, 2020). This data loss is divided by $a_{\mathrm{BS}} = 20m_{\mathrm{tr}}/n_{\mathrm{tr}}$, the expected number of observed events in a uniformly sampled stratum, before adding ridge. Here n_{tr} and m_{tr} are the training subject and event counts. The denominator is fixed from the training split rather than the realized event count of each stratum. AMSGrad uses uncorrected moments $(0.9, 0.99)$, $\epsilon = 10^{-8}$, $\eta_t = \eta_0/\sqrt{t}$ $(t \ge 1)$, and arithmetic iterate averaging. Strata are formed from shuffled training observations; incomplete strata batches are discarded at epoch boundaries.

Cox-CC. Adam minimizes the case-control loss (Kvamme et al., 2019)

$$\frac{1}{Q}\sum_{i \in \text{sampled cases}} \log\left(1 + \sum_{r=1}^{c} e^{z_{j_r} - z_i}\right)$$

plus ridge, including the case score in the denominator. Batch LSE and Cox-CC sample with replacement, allow controls equal to the case, and do not deduplicate score evaluations.

Batch LSE, Minibatch Cox, and Cox-CC use Adam with moment decay rates $(0.9, 0.999)$, $\epsilon = 10^{-8}$, bias correction, and a constant selected learning rate. These local linear implementations use Breslow ties and stable log-sum-exp. BigSurv jitter and Cox-CC score clamping/additional shrinkage are disabled.

D.3 Tuning menus and selection

Twelve candidate configurations per method use tuning seeds 100 and 101:

$$h_* = \arg\min_h \frac{1}{2} \sum_{r\in\{100,101\}} \min_{j\in\mathcal{J}_{h,r}} \ell_{\mathrm{val}}(\beta_{h,r,j}), \tag{74}$$

where $\mathcal{J}_{h,r}$ contains saved checkpoints; ties favor the lower candidate index. Each method has two structural settings and six learning rates per setting (Table 2), giving 600 configuration–seed evaluations across the five datasets and five methods. For each dataset, all five configurations are fixed before its final fits. Each selected configuration is trained from zero with optimizer seeds 0–9 on the same fixed split, giving 250 final runs in total. Per-fit caps are $U_{\mathrm{HPO}} = \max\{3 \cdot 10^6, 100N\}$ and $U_{\mathrm{final}} = 1.5U_{\mathrm{HPO}}$, where N is the retained pre-split sample size. Budgets are enforced separately for each run.

Table 2: Twelve candidate configurations per method: two structural settings and six learning rates per setting, shared across datasets. Here b is batch size, Q is the number of sampled cases, c the controls per case, and ℓ the strata per update.

Method	Candidates
Ours	$(b, \eta_0) \in \{128, 256\} \times \{0.003, 0.01, 0.03, 0.1, 0.3, 1\}$
Batch LSE	$(Q, c) = (8, 8)$: $\eta \in \{0.0009, 0.003, 0.009, 0.03, 0.09, 0.3\}$;
	$(Q, c) = (16, 64)$: $\eta \in \{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1\}$
Minibatch Cox	$(b, \eta) \in \{32, 256\} \times \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3\}$
BigSurv	Stratum size 20; $(\ell, \eta_0) \in \{1, 16\} \times \{0.006, 0.02, 0.06, 0.2, 0.6, 2\}$
Cox-CC	$Q = 32$; $(c, \eta) \in \{1, 8\} \times \{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1\}$

Table 3: Configurations minimizing the mean, over two tuning seeds, of the minimum saved validation Cox loss within the tuning budget. Tuples are (b, η_0) for Ours, (Q, c, η) for Batch LSE, (b, η) for Minibatch Cox, (ℓ, η_0) for BigSurv, and (c, η) for Cox-CC (with $Q = 32$). Candidate menus are given in Table 2.

Method	SUPPORT2	NWTCO	Correlated 100k	Correlated 1M	Independent 100k
Ours	$(256, 0.03)$	$(128, 0.03)$	$(128, 0.1)$	$(256, 0.1)$	$(128, 0.03)$
Batch LSE	$(16, 64, 0.001)$	$(16, 64, 0.01)$	$(8, 8, 0.009)$	$(16, 64, 0.0003)$	$(16, 64, 0.0003)$
Minibatch Cox	$(256, 0.01)$	$(256, 0.03)$	$(32, 0.003)$	$(256, 0.001)$	$(256, 0.001)$
BigSurv	$(1, 0.02)$	$(16, 0.06)$	$(1, 2)$	$(1, 2)$	$(16, 0.06)$
Cox-CC	$(8, 0.001)$	$(8, 0.03)$	$(8, 0.03)$	$(1, 0.0003)$	$(8, 0.0003)$

D.4 Vector-work accounting and timing

A vector unit is one length-d inner product, accumulation, arithmetic pass, or norm; $X\beta$ and $X^\top v$ each cost one per processed row. Our updates cost three batch passes (risk scores, weighted risk and event feature sums) and seven coefficient passes (gradient difference, ridge, update, norm, and averaging); active projection adds one. Setup requires no feature-vector arithmetic. Baseline Adam/AMSGrad use a fixed estimate of 12 passes per update: total $2P+15t$ for P physical scores and t updates, plus $3t$ for BigSurv averaging. Units exclude scalar nonlinearities, shifts, sampling, sorting, searches, memory copies, and runtime overhead; this work enters elapsed time when performed within the timed intervals below. Units model vector work, not hardware FLOPs.

All runs use float64 arithmetic, one compute thread, and one training worker. Our method uses NumPy and the baselines use PyTorch. Reported time is method setup plus optimizer execution,

including sampling and budget checks. Common data preparation, reference computation, checkpoint snapshot recording, full-data diagnostics, and serialization are excluded. Runtime comparisons characterize these implementations and were recorded in separate execution sessions; software and hardware details accompany the reproducibility artifact.

D.5 Metrics, checkpoints, and aggregation

For split D with m_D events,

$$\ell_D(\beta) = \frac{1}{m_D} \sum_{i \in \mathcal{E}_D} \left[\log \sum_{j \in \mathcal{R}_{i,D}} e^{x_j^\top \beta} - x_i^\top \beta \right].$$

On the training split, $\ell_{\mathrm{tr}}(\beta) + \lambda \|\beta\|^2 / 2$ differs from $\mathcal{L}(\beta)$ in (11) only by $m^{-1} \sum_{i \in \mathcal{E}} \log n_i$. We report the signed gap $\mathcal{L}(\beta) - \mathcal{L}(\beta_{\mathrm{ref}})$, with recorded resolution approximately $1.2\text{–}1.8 \cdot 10^{-13}$. References minimize the full-risk-set, event-normalized Breslow objective with $\lambda = 10^{-3}$, using float64 L-BFGS/strong-Wolfe from zero and a Newton-polishing fallback. All five stopped at full regularized gradient norm below 10^{-10}. Reference computation is separate from compared trajectories.

Validation/test use unpenalized ℓ_D and reporting-only Harrell C-index. Each tuning and final run uses 48 geometrically spaced work targets from 128 to its respective cap, together with the initial and actual terminal states. We record the first completed update crossing each target and coalesce targets crossed by the same update. Target locations are shared within a dataset and phase, while realized checkpoint costs can differ between methods. The saved checkpoint with the smallest validation Cox loss is selected, with ties resolved in favor of the earliest checkpoint. Full-data diagnostics are computed after training and do not affect updates.

A run attains a threshold through the cap at the earliest saved post-update checkpoint at which the gap plus reference resolution, and its value at every later saved checkpoint, is at most that threshold. We use thresholds $10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}$. A first crossing at the terminal checkpoint also qualifies. Costs use recorded points without interpolation; nonattainment is reported separately from the median cost among attainers. This assesses saved checkpoints rather than all iterates.

Figures show ten seed traces, medians, and 25th–75th percentile bands. Aggregation follows linear interpolation in displayed axis coordinates on common observed support, without extrapolation or best-so-far smoothing. Tables report means and sample SD where stated. Variability concerns optimizer randomness on one split, not independent cohorts or realizations.

For each dataset, all 50 validation-selected models are fixed before their test evaluation. Each selected model is scored once on its test split. These test cohorts were also evaluated during earlier exploratory work, so the predictive results are descriptive rather than a new independent confirmation.

E Additional results

E.1 Optimization accuracy and target attainment

Our method achieves the lowest median terminal full-Cox training gap on all five datasets at matched vector-work caps (Table 4). BigSurv is the closest baseline, with median gaps 3.74, 2.38, 16.14, 26.15, and 1.22 times ours in table order. Our method has a smaller gap than BigSurv for all ten seed labels on four datasets and for eight on Independent 100k. These results compare complete training procedures against a common regularized full-Cox reference; their native objectives and optimizers differ (Appendix D.2).

At the 10^{-5} target, our method is the only method to attain the required accuracy within the final caps: seven of ten runs on Correlated 100k and all ten on Correlated 1M (Table 5). At 10^{-3}, our method and BigSurv attain the target in all ten runs on every dataset. Among these two methods, our method uses less median work on Correlated 1M, while BigSurv uses less on SUPPORT2, NWTCO, Correlated 100k, and Independent 100k. No method attains 10^{-7} or 10^{-9} within these caps. Reported costs are conditional on sustained attainment at saved checkpoints through the cap and depend on the target accuracy.

Table 4: Median terminal full-Cox training gap over ten optimizer seeds

Dataset	Ours	Batch LSE	Minibatch Cox	BigSurv	Cox-CC
SUPPORT2	$\mathbf{3.93 \cdot 10^{-5}}$	$1.47 \cdot 10^{-3}$	$2.13 \cdot 10^{-3}$	$1.47 \cdot 10^{-4}$	$1.16 \cdot 10^{-3}$
NWTCO	$\mathbf{4.96 \cdot 10^{-5}}$	$7.77 \cdot 10^{-3}$	$2.98 \cdot 10^{-3}$	$1.18 \cdot 10^{-4}$	$1.19 \cdot 10^{-2}$
Correlated 100k	$\mathbf{7.92 \cdot 10^{-6}}$	$1.76 \cdot 10^{-2}$	$1.61 \cdot 10^{-3}$	$1.28 \cdot 10^{-4}$	$1.24 \cdot 10^{-2}$
Correlated 1M	$\mathbf{1.39 \cdot 10^{-6}}$	$4.79 \cdot 10^{-4}$	$7.11 \cdot 10^{-4}$	$3.64 \cdot 10^{-5}$	$6.46 \cdot 10^{-4}$
Independent 100k	$\mathbf{1.36 \cdot 10^{-5}}$	$4.49 \cdot 10^{-4}$	$2.27 \cdot 10^{-4}$	$1.65 \cdot 10^{-5}$	$2.68 \cdot 10^{-4}$

Table 5: Sustained training-gap attainment at saved post-update checkpoints through the final cap. Threshold columns give the number of attainers out of ten runs. Costs are median work in millions of vector units among attainers. A first crossing at the terminal checkpoint qualifies, a dash indicates no attainment.

Dataset	Method	10^{-3}	Cost (M)	10^{-5}	Cost (M)
SUPPORT2	Ours	10/10	0.3115	0/10	–
SUPPORT2	Batch LSE	1/10	4.4980	0/10	–
SUPPORT2	Minibatch Cox	0/10	–	0/10	–
SUPPORT2	BigSurv	10/10	0.0654	0/10	–
SUPPORT2	Cox-CC	3/10	4.4999	0/10	–
NWTCO	Ours	10/10	0.3108	0/10	–
NWTCO	Batch LSE	0/10	–	0/10	–
NWTCO	Minibatch Cox	1/10	4.4998	0/10	–
NWTCO	BigSurv	10/10	0.0474	0/10	–
NWTCO	Cox-CC	0/10	–	0/10	–
Correlated 100k	Ours	10/10	0.9767	7/10	11.7018
Correlated 100k	Batch LSE	0/10	–	0/10	–
Correlated 100k	Minibatch Cox	4/10	12.0642	0/10	–
Correlated 100k	BigSurv	10/10	0.4127	0/10	–
Correlated 100k	Cox-CC	0/10	–	0/10	–
Correlated 1M	Ours	10/10	0.9571	10/10	25.1968
Correlated 1M	Batch LSE	10/10	61.4778	0/10	–
Correlated 1M	Minibatch Cox	6/10	72.6711	0/10	–
Correlated 1M	BigSurv	10/10	1.2885	0/10	–
Correlated 1M	Cox-CC	9/10	13.9025	0/10	–
Independent 100k	Ours	10/10	0.3621	0/10	–
Independent 100k	Batch LSE	10/10	5.5559	0/10	–
Independent 100k	Minibatch Cox	10/10	0.2205	0/10	–
Independent 100k	BigSurv	10/10	0.0388	0/10	–
Independent 100k	Cox-CC	10/10	1.2523	0/10	–

E.2 PREDICTIVE PERFORMANCE

Tables 6 and 7 report validation-selected metrics. Mean test Cox loss is lowest for our method on NWTCO, Correlated 100k, and Independent 100k, for Minibatch Cox on SUPPORT2, and for BigSurv on Correlated 1M. Mean test C-index is highest for our method on NWTCO and Correlated 100k, for BigSurv on SUPPORT2 and Independent 100k, and for Cox-CC on Correlated 1M.

Table 6: Test metrics at validation-selected checkpoints: mean ± sample SD over ten optimizer seeds. One fixed split; unpenalized Cox loss per event (↓), Harrell C-index (↑).

Dataset	Method	Test Cox loss	Test C-index
SUPPORT2	Ours	6.866840 ± 0.000700	0.602019 ± 0.000693
SUPPORT2	Batch LSE	6.867794 ± 0.001330	0.600758 ± 0.001080
SUPPORT2	Minibatch Cox	6.866488 ± 0.002054	0.601296 ± 0.001637
SUPPORT2	BigSurv	6.867411 ± 0.000304	0.602970 ± 0.000435
SUPPORT2	Cox-CC	6.867097 ± 0.001045	0.601770 ± 0.001126
NWTCO	Ours	6.227649 ± 0.001645	0.731458 ± 0.000142
NWTCO	Batch LSE	6.241507 ± 0.015734	0.726018 ± 0.007678
NWTCO	Minibatch Cox	6.248697 ± 0.012824	0.723723 ± 0.004764
NWTCO	BigSurv	6.232977 ± 0.005778	0.730110 ± 0.001328
NWTCO	Cox-CC	6.263264 ± 0.020098	0.717591 ± 0.008785
Correlated 100k	Ours	8.420984 ± 0.000132	0.635188 ± 0.000056
Correlated 100k	Batch LSE	8.424820 ± 0.001777	0.634273 ± 0.001613
Correlated 100k	Minibatch Cox	8.422421 ± 0.001239	0.634502 ± 0.000873
Correlated 100k	BigSurv	8.421654 ± 0.000315	0.634838 ± 0.000220
Correlated 100k	Cox-CC	8.422216 ± 0.001405	0.634370 ± 0.000942
Correlated 1M	Ours	10.702666 ± 0.000042	0.639196 ± 0.000023
Correlated 1M	Batch LSE	10.702925 ± 0.000225	0.639051 ± 0.000096
Correlated 1M	Minibatch Cox	10.702853 ± 0.000179	0.639099 ± 0.000066
Correlated 1M	BigSurv	10.702654 ± 0.000035	0.639178 ± 0.000019
Correlated 1M	Cox-CC	10.702712 ± 0.000103	0.639204 ± 0.000037
Independent 100k	Ours	8.725094 ± 0.000077	0.630792 ± 0.000043
Independent 100k	Batch LSE	8.725489 ± 0.000289	0.630621 ± 0.000188
Independent 100k	Minibatch Cox	8.725183 ± 0.000084	0.630782 ± 0.000090
Independent 100k	BigSurv	8.725162 ± 0.000225	0.630811 ± 0.000022
Independent 100k	Cox-CC	8.725228 ± 0.000211	0.630728 ± 0.000135

Table 7: Validation metrics at checkpoints selected by validation Cox loss: mean ± sample SD over ten optimizer seeds. One fixed split; unpenalized Cox loss per event (↓), Harrell C-index (↑).

Dataset	Method	Validation Cox loss	Validation C-index
SUPPORT2	Ours	6.871160 ± 0.000429	0.594114 ± 0.000440
SUPPORT2	Batch LSE	6.871841 ± 0.000956	0.593927 ± 0.001286
SUPPORT2	Minibatch Cox	6.870807 ± 0.000519	0.594091 ± 0.000511
SUPPORT2	BigSurv	6.871334 ± 0.000152	0.594797 ± 0.000370
SUPPORT2	Cox-CC	6.870968 ± 0.000646	0.594495 ± 0.000652
NWTCO	Ours	6.153779 ± 0.000821	0.723944 ± 0.001528
NWTCO	Batch LSE	6.141673 ± 0.003232	0.728380 ± 0.002980
NWTCO	Minibatch Cox	6.139356 ± 0.005738	0.728560 ± 0.003219
NWTCO	BigSurv	6.141284 ± 0.005943	0.727106 ± 0.001478
NWTCO	Cox-CC	6.135449 ± 0.005278	0.731840 ± 0.002403
Correlated 100k	Ours	8.402429 ± 0.000093	0.642029 ± 0.000081
Correlated 100k	Batch LSE	8.404346 ± 0.001473	0.641605 ± 0.001088
Correlated 100k	Minibatch Cox	8.402208 ± 0.000598	0.641995 ± 0.000268
Correlated 100k	BigSurv	8.402015 ± 0.000557	0.642109 ± 0.000269
Correlated 100k	Cox-CC	8.402560 ± 0.000860	0.641725 ± 0.000603
Correlated 1M	Ours	10.705947 ± 0.000033	0.637771 ± 0.000027
Correlated 1M	Batch LSE	10.706178 ± 0.000166	0.637626 ± 0.000096
Correlated 1M	Minibatch Cox	10.706057 ± 0.000099	0.637729 ± 0.000065
Correlated 1M	BigSurv	10.705949 ± 0.000016	0.637750 ± 0.000011
Correlated 1M	Cox-CC	10.705979 ± 0.000040	0.637758 ± 0.000047
Independent 100k	Ours	8.711592 ± 0.000117	0.634014 ± 0.000091
Independent 100k	Batch LSE	8.711744 ± 0.000291	0.633971 ± 0.000185
Independent 100k	Minibatch Cox	8.711506 ± 0.000052	0.634057 ± 0.000052
Independent 100k	BigSurv	8.711569 ± 0.000096	0.634021 ± 0.000090
Independent 100k	Cox-CC	8.711505 ± 0.000148	0.634071 ± 0.000087

F APPROXIMATION DIAGNOSTICS

On each training dataset we evaluate $v \in \{0, 0.5\beta_{\text{ref}}, \beta_{\text{ref}}, 1.5\beta_{\text{ref}}\}$, where β_{ref} is the saved regularized full-Cox reference. The grouping grid is $\delta \in \{0.2, 0.1, 0.05, 0.025, 0.0125\}$; the softplus grid is $\rho \in \{10^{-2}, 3 \cdot 10^{-3}, 10^{-3}, 3 \cdot 10^{-4}, 10^{-4}, 3 \cdot 10^{-5}\}$ at $\delta = 0.05$.

We use $\mathcal{L}$, $\widetilde{\mathcal{L}}_\delta$, and $\mathcal{L}_{\rho,\delta}$ from Section 4, all with $\lambda = 10^{-3}$. The grouped objective uses exact group normalizers; the softplus objective profiles the shifts over $\mathbb{R}^K$. The errors are Euclidean coefficient-gradient differences:

$$e_{\text{group}}(v, \delta) = \|\nabla\widetilde{\mathcal{L}}_\delta(v) - \nabla\mathcal{L}(v)\|, \tag{75}$$

$$e_{\text{soft}}(v, \rho) = \|\nabla\mathcal{L}_{\rho,0.05}(v) - \nabla\widetilde{\mathcal{L}}_{0.05}(v)\|. \tag{76}$$

Softplus shifts are profiled by safeguarded scalar roots with tolerance $5 \cdot 10^{-13}$. All 100 grouping and 120 softplus comparisons are retained; near-zero errors at $v = 0$ are excluded from logarithmic fits.

Figure 2 and Table 8 use every grid point at β_{ref}. Ordinary least-squares log–log slopes are 1.37–1.87 for grouping and 0.997–1.000 for softplus error. Softplus error is nearly linear in ρ over the tested range. Grouping changes discretely as δ varies, so its finite-grid slopes also reflect changes in the partition. The theoretical $O(\delta^2)$ bound applies under the stated compression conditions and does not prescribe an exact finite-grid slope. These diagnostics characterize approximation at fixed coefficients; they do not estimate convergence over optimizer iterations.

Table 8: Group count K at $\delta = 0.05$ and ordinary least-squares log–log slopes of coefficient-gradient error against δ (grouping) and ρ (softplus), using every point in the respective grid at β_{ref}.

Dataset	K	δ slope	ρ slope
SUPPORT2	61	1.783	1.000
NWTCO	15	1.601	0.997
Correlated 100k	133	1.372	0.999
Correlated 1M	178	1.716	0.999
Independent 100k	166	1.866	0.999

The markers evaluate $r_\delta \le 1 - q$, $L_q(v)r_\delta \le 1/2$, and $\rho\kappa(v) \le 1/9$, with $q = 0.5$ and $r_\delta = \delta/(1 + \delta)$. Here $L_q(v)$ and $\kappa(v)$ are the expressions in (19) and (18) evaluated at $\beta = v$ before taking the supremum over $\mathcal{B}$. The applicable conditions hold in 80/100 grouping and 96/120 softplus comparisons. These pointwise checks do not certify the assumptions uniformly over the radius-10 ball. The calculations preserve event multiplicity and the Breslow convention for ties.

G COMPLETE TRAJECTORY FIGURES

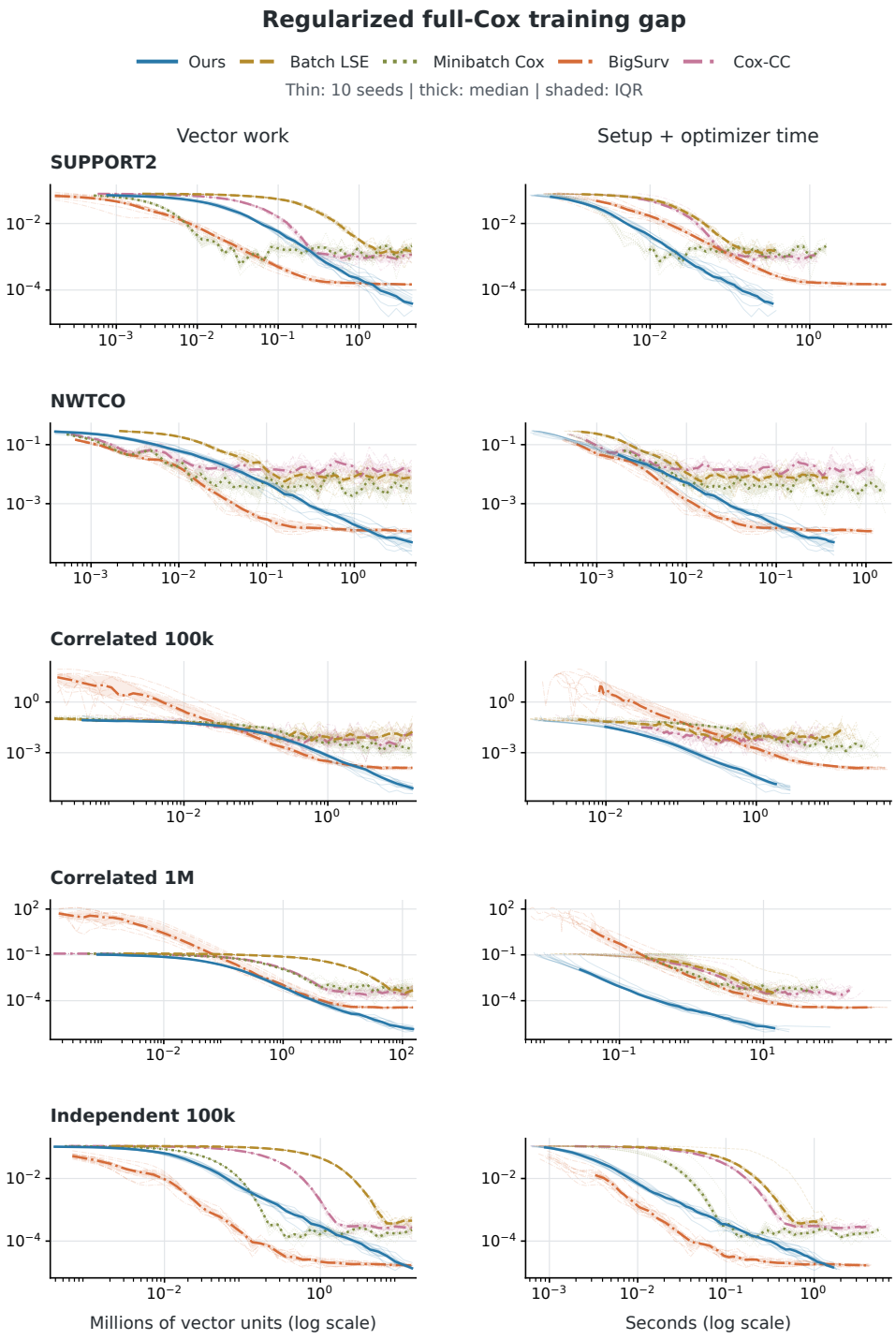


Figure 3: Regularized full-Cox training gap against vector work and setup plus optimizer time. Thin lines show individual optimizer seeds, thick lines and bands give the median and interquartile range. Ours and BigSurv use running arithmetic averages of post-update coefficient iterates; Ours averages after projection. The other methods use current coefficient iterates.

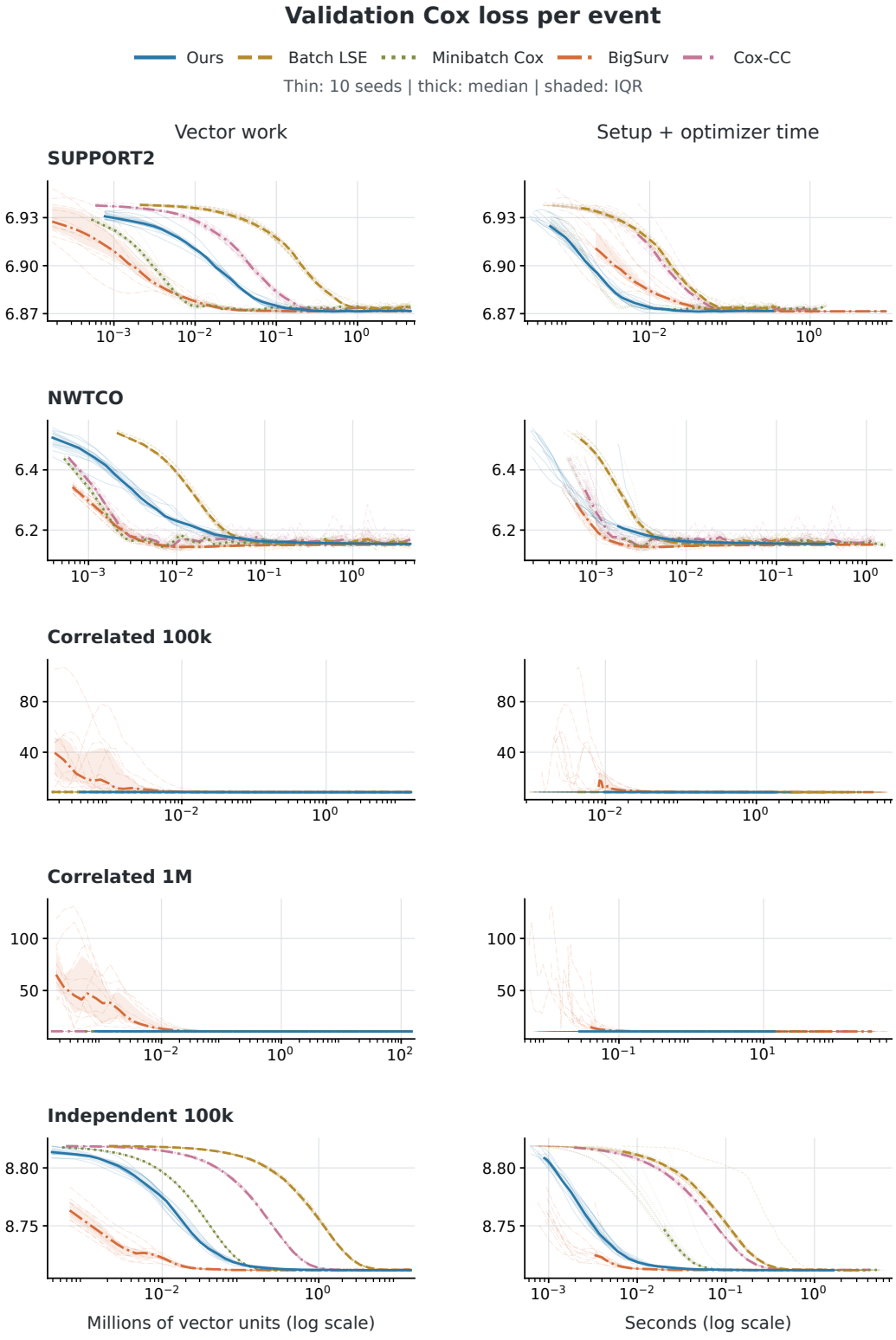


Figure 4: Unpenalized validation Cox loss per event against vector work (left) and setup plus optimizer time (right); lower is better. Matched panels share Y-axis limits.

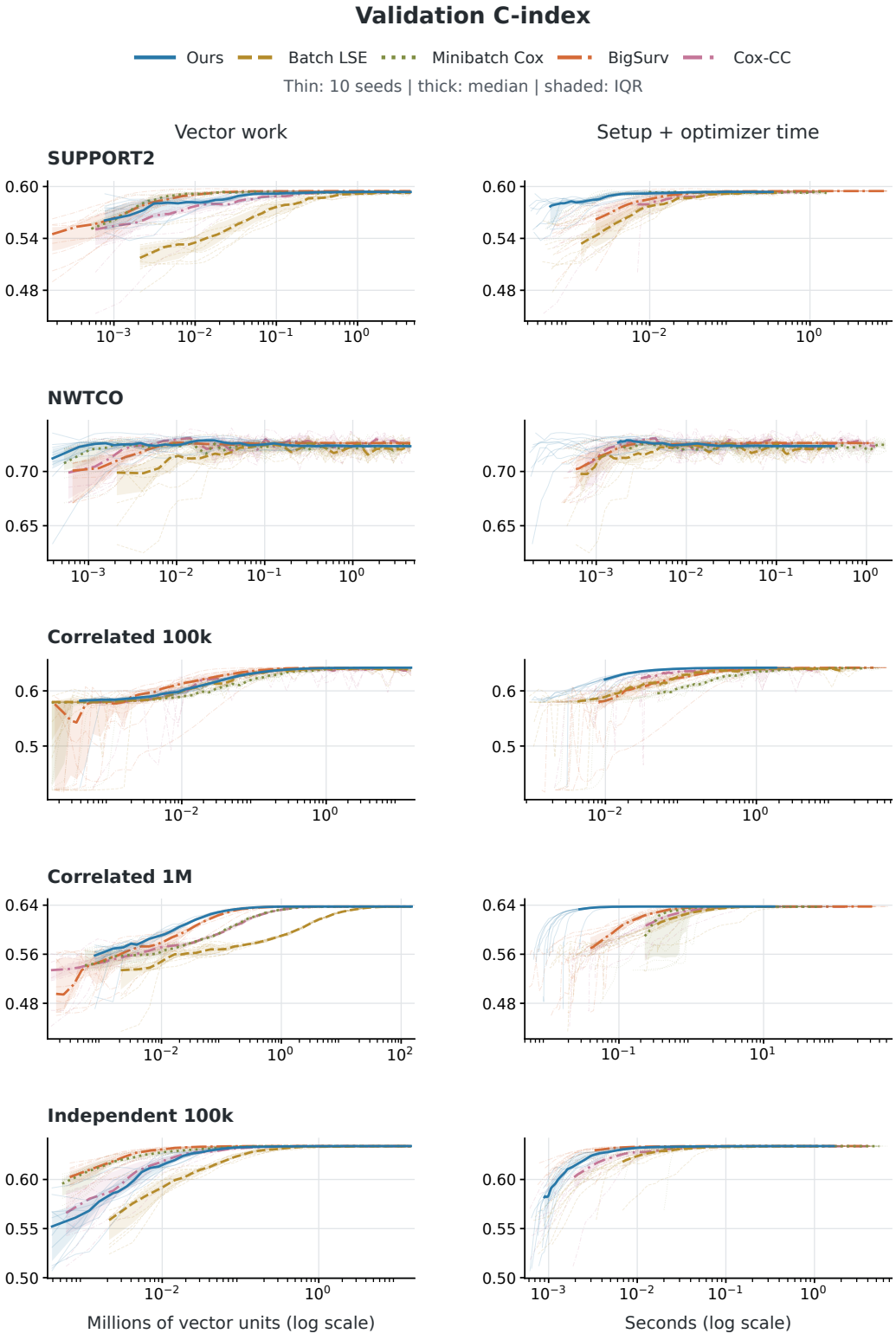


Figure 5: Validation Harrell C-index against vector work (left) and setup plus optimizer time (right); higher is better. Matched panels share Y-axis limits.